ESSAYS IN LABOR ECONOMICS

A Dissertation

Presented to the Faculty of the Graduate School

of Cornell University

in Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy

by

Andrew Shaw Green

May 2017

ESSAYS IN LABOR ECONOMICS

Andrew Shaw Green, Ph.D.

Cornell University 2017

This dissertation contributes to the understanding of employer-employee bargaining over hours of work, and the use of administrative data to better understand labor market statistics.

In chapter 1, "Hours Off the Clock," I address a simple but confounding research question: To what extent do workers work more hours than they are paid for? The relationship between hours worked and hours paid, and the conditions under which employers can demand more hours "off the clock," is not well understood. The answer to this question affects worker welfare, as well as wage and hour regulation. In addition, work off the clock has important implications for the measurement and cyclical movement of productivity and wages. In this chapter, I construct a unique administrative dataset of hours paid by employers linked to a survey of workers on their reported hours worked to measure work off the clock. Using cross-sectional variation in local labor markets, I find only a small cyclical component to work off the clock. The results point to labor hoarding rather than efficiency wage theory, indicating work off the clock cannot explain the counter-cyclical movement of productivity. I find workers employed by small firms, and in industries with a high rate of wage and hour violations are associated with larger differences in hours worked than hours paid. These findings suggest the importance of tracking hours of work for enforcement of labor regulations.

In chapter 2, "Hours Adjustments: Evidence from Linked Employer-Employee Data," coauthored with fellow graduate student Nellie Zhao, we provide the first look at administrative data on hours worked within firms. We document the extent to which part-time work varies across industries, and confirm that part-time work is concentrated in rela-

tively low-wage service sectors. Further, we take advantage of the longitudinal nature of our dataset and analyze the prevalence of transitions between part-time and full-time work within the same job. We show that the share of new full-time or part-time jobs that are created due to within-job hours changes varies greatly across industries.

In chapter 3, "Total Error and Variability Measures with Integrated Disclosure Limitation for Quarterly Workforce Indicators and LEHD Origin Destination Employment Statistics in OnTheMap," coauthored with Kevin L. McKinney, Lars Vilhuber, and John M. Abowd, we report results from the first comprehensive total quality evaluation of five major indicators in the U.S. Census Bureau's Longitudinal Employer-Household Dynamics (LEHD) Program Quarterly Workforce Indicators (QWI). We conducted the evaluation by generating multiple threads of the edit and imputation models used in the LEHD Infrastructure File System. These threads conform to the Rubin (1987) multiple imputation model, with each thread or implicate being the output of formal probability models that address coverage, edit, and imputation errors. Design-based sampling variability and finite population corrections are also included in the evaluation. We derive special formulas for the Rubin total variability and its components that are consistent with the disclosure avoidance system used for QWI. These formulas allow us to publish the complete set of detailed total quality measures for QWI and LODES. The analysis reveals that the five publication variables are estimated very accurately for tabulations involving at least 10 jobs. Tabulations involving three to nine jobs have acceptable quality. Tabulations involving zero, one or two jobs have substantial total variability.

BIOGRAPHICAL SKETCH

Andrew Shaw Green received his B.A. from the University of Rochester in 2008. Starting in 2009 he worked as a research assistant of labor economics at the Economic Policy Institute, an economics think tank based in Washington, D.C. In 2011, he began graduate studies in economics at Cornell Univeristy. After three years based in Ithaca, New York, Andrew completed his PhD while working as an intern at the U.S. Census Bureau in Washington, D.C.

ACKNOWLEDGEMENTS

I would never have seen this dissertation to completion without the support of my family. My wife, Sasha Kapadia, provided invaluable support navigating life during graduate school. She is unique and talented in countless ways. Without her I would never have reached this stage. My mother, Kathy Shaw, provided tremendous love and support but also countless advice on the research process. She is the main reason I am who I am today. To Benjamin, Rebecca, and Eric, I am forever indebted for your love and support.

I am no less grateful to my entire dissertation committee for their thoughtful guidance. Special thanks and gratitude is due to the committee chair John Abowd whose comments, suggestions, and edits saw my first research questions evolve into a completed dissertation. A close second to his academic advising, John's mentoring in all aspects of the economics profession, and his flexibility and understanding with regard to the many challenges of completing a dissertation have been invaluable. Lars Vilhuber, a committee member, provided thoughtful advice and comments, as well as immeasurable support for using administrative data. This thesis would not be possible without him. Special acknowledgement is also due to committee members Karel Mertens and Richard Mansfield, whose comments and suggestions have greatly improved the quality of this work.

I completed the final three years of this dissertation at the Center for Economic Studies in the U.S. Census Bureau. There are too many wonderful economists at the Census Bureau to acknowledge in this space. Their comments and support were above the call of duty, and to them I owe much thanks. Two people in particular deserve special mention. Mark Kutzbach provided helpful comments on many drafts and handled the disclosure for chapter 1. For his time and patience, I am forever indebted. Special thanks is also due to Erika McEntarfer for her generosity and support supervising me in Washington, D.C.

This dissertation benefited greatly from the comments and suggestions of many friends and fellow graduate students. First and foremost are coauthors Nellie Zhao and

Kevin McKinney. Nellie and Kevin provided my first introduction to the Census Bureau's data and computing environment. I have grown and learned immensely from writing and collaborating with such fine economists. I also owe a special thanks to fellow graduate students Tamara McGavock, Hautahi Kingi, and Kathryn Edwards for their time and patience reading various drafts.

Finally, I want to acknowledge the contributions of various Census Bureau employees who contributed to Appendix C.1. Portions of this appendix are based on an unpublished technical memo dated February 1, 2011 by John Abowd, Henry Hyatt, Mark Kutzbach, Erika McEntarfer, Kevin McKinney, Michael Strain, Lars Vilhuber, and Chen Zhao. Chapter 3 would not have been possible without the support of NSF Grants SES-0922005, BCS 0941226, TC-1012593, and SES-1131848. The research in this dissertation uses data from the U.S. Census Bureau's Longitudinal Employer-Household Dynamics Program, which was partially supported by the following National Science Foundation Grants: SES-9978093, SES-0339191 and ITR-0427889; National Institute on Aging Grant AG018854; and grants from the Alfred P. Sloan Foundation.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

HOURS OFF THE CLOCK

1.1 Introduction

How many hours do people work? Myriad government surveys of hours worked from households, and hours paid from establishments exist to answer this question, though each has drawbacks.[1] How much time employees spend at work, and whether this time is explicitly tracked and bargained for, is neither well measured nor understood. In addition to worker welfare, the difference between hours paid and hours worked has implications for the cyclical movement of productivity and wages. Specifically, off-the-clock work[2] is a possible explanation for the change in productivity from pro-cyclical to counter-cyclical during the last three business cycles.

The difficulty measuring work off the clock is not only a concern for economists and government statistical agencies, it is also essential for the understanding of firm profits and worker welfare. In the last few years, stories in the popular press recount hourly workers asked to show up to work and told to wait – without pay – until demand picked up.[3] Other times wage theft was more explicit, with employers doctoring reports of hours worked to show a higher hourly wage,[4] or workers asked to keep working after they had clocked out.[5] Salaried workers, too, were frequently asked to pick up some or all of the work for colleagues who were laid off.[6] The various stories are not uniform to all

[1] Establishment surveys of hours paid likely miss long hours for salaried workers and off-the-clock work. Household surveys rely on accurate recall of all jobs.

[2] I will refer to "off-the-clock" work synonymously with "the difference between hours worked and hours paid". I use this phrase only for its brevity.

[3] "More Workers Are Claiming 'Wage Theft'." *The New York Times*, Aug. 31, 2014.

[4] "Squeezed garment factories use check cashing services to mask true wages, workers say." *The Los Angeles Times*, Jul. 30, 2016.

[5] "Nearly 10,000 Chipotle Workers Join Class Action Wage Lawsuit." *The New York Daily News*, Aug. 30, 2016.

[6] "All Work and No Pay: The Great Speedup." *Mother Jones*, July/August, 2011.

workers, but they all point to the various dynamics that influence bargaining over time in the workplace and how much work time happens off the clock. Empirical research that documents off-the-clock work specifically, and labor compliance more broadly, is growing, but little research using representative government datasets exists.[7]

In this study, I examine how shocks to labor demand and firm characteristics influence work off the clock. I construct a unique dataset of survey responses of hours worked from the U.S. Census Bureau's American Community Survey (ACS) with administrative data on hours paid from the U.S. Census Bureau's Longitudinal Employer-Household Dynamics (LEHD) program. I adjust the ACS survey weights to account for the shift in frame, so the analysis sample remains representative. To the best of my knowledge, this is the first study that links hours worked to hours paid at the *person*-level.

The dataset allows me to first answer an elementary, but essential question: how much do people work? The unconditional means of the analysis sample find that annual hours paid for full-year workers is 1,946 [500.3] compared to 2,079 [504.8] hours worked.[8] The difference of a little more than 130 hours per year works out to roughly an extra three weeks per year assuming the standard 40 hour work week. The unconditional mean log difference is 0.079 [0.237], which is close to the log difference of means. The unconditional means mask significant heterogeneity in differences between subgroups. In particular, workers who self-report working less than 40 hours or exactly 40 hours per week – the standard workweek in the U.S. – have a mean log difference between hours worked and hours paid of 0.031 [0.268] and 0.032 [0.179], respectively. For those who self-report working more than 40 hours per week, the mean log difference is 0.210 [0.238]. These are the first results to confirm that firms poorly track the hours of workers who work more than the standard workweek.

[7] Bernhardt et al. (2013) and Milkman et al. (2012) are recent examples.
[8] All standard errors are in parentheses. All standard deviations are in brackets.

After quantifying the extent of off-the-clock work, I use variation in local labor market conditions to test whether and how shocks to labor demand effect off-the-clock work. I regress the log difference on the unemployment rate in the commuting zone of the ACS respondent at the time of her interview. The coefficient estimate for the effect of local labor market conditions is −0.00191 (0.00077), indicating that tighter labor markets increase off-the-clock work. A one percentage point decrease in the unemployment rate increases off-the-clock work by 0.19%, or an extra 4 hours annually. Further analysis reveals that the effect is driven by workers in production and non-supervisory occupations, low-skilled workers, as well as workers likely paid by the hour.

Due to concerns about about the endogeneity of hours reporting and other labor market programs that may effect labor force participation, I instrument the unemployment rate using a shift-share predicted employment index (Bartik, 1991). The instrumental variable coefficient estimates are slightly larger in magnitude, though still relatively small, with a coefficient estimate of −0.00274 (0.00154). A negative estimate indicates that off-the-clock work is not a viable explanation for the changing cyclicality of productivity.

Economic theory gives us insight as to why hours paid and hours worked should diverge. The negative coefficient estimate points to firms engaging in labor hoarding. In labor hoarding models, firms hold labor in excess of production requirements during a drop in demand. This is usually attributed to costs of adjusting employment, such as the difficulty of training new workers when demand picks up. In accordance with my results, labor hoarding models find relatively sluggish employment adjustment in response to a shock, with firms using the intensive margin to adjust labor inputs. One important implication is that productivity is pro-cyclical.

In light of the result that off-the-clock work is probably pro-cyclical and driven by low-skill workers, I test for explanations centered on labor compliance. Although still an emerging literature, research finds that smaller firms are much less likely to comply with

labor regulations. Consistent with this literature, I find that firms in the smallest firm size category, 0-19 employees, report higher incidence of work off the clock compared to firms with greater than 2,500 employees with a log difference of 0.0231 (0.0053). I find the effect is driven by production and non-supervisory workers who are likely paid on an hourly basis. I also show that off-the-clock work is concentrated in industries where wage and hour violations are prevalent.

The results on the cyclical nature of off-the-clock work challenge a recent literature testing for efficiency wage explanations for the counter-cyclical nature of productivity. Lazear et al. (2015) and Burda et al. (2016) use local labor market variation to test for greater effort in slack labor markets. Unlike this paper, they do not look for work off the clock, rather they model greater effort per unit of time at work. The empirical strategy in Lazear et al. (2015) employs data on a single firm with a wide geographic dispersion of establishments, that tracks piece-rate production. Burda et al. (2016) use an empirical strategy similar to Lazear et al. (2015), but they use the American Time Use Survey to measure time at work actually working, which is a proxy for greater effort. In contrast to my paper, both studies find evidence that greater local labor market slack is associated with greater effort provision.

My estimates quantifying off-the-clock work and its implications for productivity statistics confirm several previous papers. Aaronson and Figura (2010) also attempt to use off-the-clock work to explain the counter-cyclical turn in productivity. They construct time series of hours worked and hours paid from the Current Population Survey and the Current Establishment Statistics, respectively. Although they must rely on aggregate data, they too find little evidence for off-the-clock work biasing productivity estimates. Eldridge and Pabilonia (2010) use the American Time-Use Survey (ATUS) and the Work Schedules and Work at Home Supplement to the Current Population Survey to address whether the incidence of working from home biases productivity statistics. They find

over the time span of their sample that unpaid work at home sometimes overstates, and other times understates the hours levels used in the BLS productivity series. The bias in all cases is exceedingly small, and unlikely to bias productivity statistics.

Recent research uses survey and administrative data to document non-compliance with minimum wage laws, overtime regulations, and work off the clock. Bernhardt et al. (2013) and Milkman et al. (2012) use the 2008 Unregulated Worker Survey, and they find that job and employer characteristics are responsible for much of the variation in non-compliance. Ji and Weil (2015) use a unique dataset of franchisor- and franchisee-owned establishments matched to Wage and Hour Administration investigations. They find that franchisee-owned establishments are more likely to commit wage and hour violations. My estimates of off-the-clock work are broadly consistent with this literature. Off-the-clock work is most concentrated in small firms, and within industries that disproportionately employ low-wage workers.

Since applied economics is moving towards the use of large administrative datasets, the results of my study also suggest caution when using administrative data to measure hours. Few studies have explicitly compared employer and employee reports of hours worked.[9] One exception is Mellow and Sider (1983) who use an employer validation supplement to the CPS to glean employer and worker responses to myriad questions. For hours, they find worker reports exceed employer reports by 3.9%, which is substantially less than the 7.9% in the preferred specification. The difference is likely due to the analysis sample in this study, which only considers full-time, full-year workers. Lastly, this study contributes to the growing body of research which uses administrative data to validate survey data.[10]

[9]See Duncan and Hill (1985), Bound and Krueger (1991), Bound et al. (1994), and Bound et al. (2001) for an overview.

[10]See Abraham et al. (2013) and Abowd and Stinson (2013).

1.2 Hours Divergence: Theory and Implications

I construct a unique dataset of hours worked and hours paid in order to try to infer the causes, incidence, and implications of off-the-clock work. A natural question quickly arises: why should we expect the difference between hours paid and hours worked to reflect anything but errors in reporting? Although measurement error is no doubt present, this section lays out established economic theories of why hours paid may diverge from hours worked. The first two theories provide opposing predictions for the movement of off-the-clock work over the business cycle. Efficiency wage models predict workers will exert more effort – greater hours worked compared to hours paid – when labor markets are slack. In contrast, theories of labor hoarding predict the opposite relationship between off-the-clock work and macroeconomic conditions.

In addition to cyclical theories of why hours paid may diverge from hours worked, I view the difference through an older literature on labor regulation compliance. In these models the firm's profit motive leads them to skirt labor laws to realize greater profits. Firms must weigh the higher profit of non-compliance against the probability of getting caught and the penalties of non-compliance. The Great Recession and the ensuing debate about the declining wages and working conditions of low wage workers have brought this topic to greater prominence in the media. Better administrative data on employment, firms, and greater transparency and data around compliance investigations have invigorated this literature.

1.2.1 Efficiency Wages & Labor Hoarding

In efficiency wage theories of the labor market, workers would like to avoid being laid off, and firms would like workers to exert effort. At least since Kalecki (1943), who noted

"under a regime of permanent full employment, the 'sack' would cease to play its role as a disciplinary measure," economists have studied the relationship between the labor market and worker effort. More recent discussions of efficiency wage models pick up with Shapiro and Stiglitz (1984). In their model, the firm's production depends on worker effort, which firms cannot perfectly monitor. Workers would prefer to shirk rather than exert effort. Firms offer wages in excess of the market clearing rate in order to induce effort. The theory provides a succinct explanation of involuntary unemployment.

The model relevant for the empirical tests in this paper does not provide a theory of unemployment. Although similar, the key is the cost to firms of replacing workers, or alternatively the cost to workers of finding a new job.[11] The driving variable is the tightness of the labor market. As the labor market becomes more slack, the cost of job loss increases due to worse prospects of finding a new job. Workers exert extra effort in the form of more hours worked in order to signal their worth to employers and avoid a lay off. In the case of hourly workers, this may be explicit off-the-clock work. For salaried workers, these extra hours are in excess of what is "normal" under more favorable labor market conditions.

In contrast, labor hoarding theories predict that labor productivity should be pro-cyclical Popular in the 1960s,[12] theories of labor hoarding hold that firms retain more workers in a downturn than production explicitly requires.[13] Firms may have incurred the costs of training workers to their specific production technology, or highly skilled workers may be scarce. In both cases firms would rather not risk laying off workers who may be difficult to rehire, or pay the upfront cost to train new hires. To meet reduced production targets, firms then adjust hours worked in order to meet production targets. If firms choose to keep a worker's labor earnings constant, hours paid may exceed hours

[11] Rebitzer (1987) is the most relevant paper capturing the former case. See Appendix A.1 for the latter case.

[12] See Oi (1962) & Fair (1969).

[13] Biddle (2014) provides a nice history of the literature.

worked. The implication for labor hoarding theory is that productivity will decline during downturns as hours paid stays relatively constant and production declines.[14]

More modern theories of labor demand would interpret labor hoarding through the lens of adjustment costs. Firms face an explicit cost to adjusting their labor on the extensive margin. Depending on the size and functional form of the costs, firms will not always adjust employment to its optimal level in response to a shock. Firms will adjust employment less than in the absence of adjustment costs and use hours to adjust total labor input to its optimal level.[15]

Efficiency wage and labor hoarding models are not mutually exclusive. In fact, both are likely present in any given employment relationship. The empirical approach employed in this study will not be able to separately identify the two. The empirical analysis in this paper serves to answer the question of which is more salient for interpreting the cyclical changes in productivity and real wages. It should therefore help guide macroeconomists on how best to incorporate the costs of separations to workers and firms in their models.

1.2.2 Labor Regulation Compliance

In addition to the cyclical forces, explicit failure to comply with labor regulations is another reason why hours worked may exceed hours paid. The Fair Labor Standards Act (FLSA), enacted in 1938, established the federal minimum wage, and effectively enshrined the 40-hour work week by requiring overtime pay of time-and-a-half for all hours worked over 40 in a week. The literature on compliance with the FLSA centers on the

[14]Employer reports of hours paid are the main input to the BLS productivity series, while hours worked is the variable of economic interest. See Eldridge and Pabilonia (2010).

[15]See Cooper et al. (2007), Caballero et al. (1997), and Hamermesh (1989a) for more recent examples. Appendix A.1 describes the theory in more detail.

firm. Firms weigh their profits from compliance against their expected profits from non-compliance (Ashenfelter and Smith, 1979). The model leads to the conclusion that firms need to take into account the costs of noncompliance, the odds of getting caught, the elasticity of labor demand, and the spread between the prevailing wage and the minimum wage in a given industry.[16]

Recent empirical research finds firm and job characteristics such as firm size, industry, and non-hourly pay arrangements drive non-compliance with labor regulations. Bernhardt et al. (2013) conduct a survey of low-wage workers in major American metropolitan areas. They survey non-compliance, but also collect detailed worker and firm characteristics. They have two important findings. First, larger firms (greater than 100 employees) are less likely to commit labor violations compared to firms less than 100 employees. Second, they find that non-hourly workers are more likely to incur wage and hour violations, and that off-the-clock work is more prevalent than straight minimum wage violations.

There are a few reasons why smaller firms may be more likely to commit labor violations. First, small firms have fewer establishments and therefore stand less of a chance of getting caught for noncompliance if enforcement is equally probable for all establishments. Second, small firms are less likely to have in-house expertise (human resource departments) to negotiate regulations (Mendeloff et al., 2006), and are less likely to be unionized.[17] Small firms tend to have less capital and rely more heavily on labor inputs. Thus, if they are going to cut costs, it will likely be on labor.[18] Finally, the fact that small firms have less capital reduces the costs of non-compliance. Firms that owe back wages have the opportunity to declare bankruptcy and forsake owed back wages. The less capital a firm has, the smaller the costs to bankruptcy.[19]

[16] See also Chang and Ehrlich (1985), and Basu et al. (2010).
[17] Weil (1991) shows labor unions correlate with OHSA investigations.
[18] See Ji and Weil (2015) for a similar discussion on franchisee vs. franchisor labor compliance.
[19] "Few California workers win back pay in wage-theft cases." *The Los Angeles Times*, April 6, 2015.

The measurement of hours is often an important determinant for wage and hour violations. Recent changes in technology and the organization of firms make measuring hours a challenge in wage and hour compliance. Tracking hours is important for assessing off-the-clock work, overtime violations, and many minimum wage violations when workers are not paid by the hour. New technology makes assessing hours more difficult as more work takes place at home with computers outside of normal business hours. In addition, tracking hours for workers who may work at multiple work sites and who may be employed by third party entities pose new challenges to enforcement agencies (Weil, 2010). In short, the measurement of hours is invaluable for effective enforcement of wage and hour violations, and it constitutes a significant margin through which many violations take place.

1.3 Data

Making inferences about the differences between hours paid and hours worked requires data on both variables. Previous estimates of the divergence between hours paid and hours worked relied on aggregated time series data. An innovation of this paper is to link the two variables at the *person* level. To the best of my knowledge, this is the first paper to explicitly link workers' reports of hours worked to employers' reports of hours paid for a representative sample of workers. The result is a survey response of hours worked and the corresponding administrative reports of hours paid from the survey respondent's employers. To construct this difference measure, I use administrative data of hours paid from the U.S. Census Bureau's Longitudinal Employer-Household Dynamics (LEHD) program linked to survey responses to the American Community Survey from the U.S. Census Bureau.

The LEHD is an administrative file system of linked employer-employee data derived

from state unemployment insurance systems. The data result from a unique partnership between states and the U.S. Census Bureau, where the states provide the Census Bureau quarterly extracts of earnings records from their unemployment insurance systems. The core file is a job-based frame, named the Employment History File (EHF), with a unique record represented by a person-firm-year-quarter link with any positive earnings in a given quarter, which covers approximately 95% of all jobs in the United States.[20] The fact that the LEHD comprises a near-universe of jobs and employer-employee links is important as it lets me account for hours paid in all jobs of the survey respondent, as well as providing the link to survey data for hours worked.

In addition to quarterly earnings, four states provide quarterly reports of hours paid. The states are Washington, Minnesota, Rhode Island, and Oregon.[21] The quarterly hours data allow me to construct a measure of total hours paid across all jobs in the previous year for each person. The final LEHD sample consists of a person-level measure of all jobs paid in the previous year for each quarter from 2010 to 2013 for these states.[22]

Data on hours worked come from the U.S. Census Bureau's American Community Survey (ACS). The American Community Survey is a rolling monthly survey of 3.5 million households each year. The ACS replaced the Census long form after the 2000 census and as such, it asks questions on housing, demographic, and economic topics. I focus on the questions on weeks paid in the past year and the usual weekly hours worked. When combined, the two variables allow me to construct a measure of usual hours worked in the past year across all jobs from the perspective of the employee.

[20] Self-employed workers are not currently incorporated into the LEHD. For a full description of the LEHD infrastructure files see Abowd et al. (2009).

[21] There does not appear to be an explicit administrative reason why some states collect hours paid in addition to quarterly earnings.

[22] The states vary considerably with respect to the time of reporting. Internal rules of the LEHD program dictate that at least three states must be used for any released results, which limits the analysis to begin in 2009 when Rhode Island first begins reporting hours until 2014 quarter one, which is the most recent hours data available for all states.

A point of clarification is needed regarding paid weeks worked for the annual hours worked measure. The ACS hours worked measure conflates both hours paid and hours worked due to the weeks worked variable including paid leave. I construct annual hours worked as the product of usual weekly hours and weeks worked. The ACS measure of annual hours worked includes some weeks for which the worker was paid, but for which no work was done. I do not adjust for this in what follows because both LEHD and ACS include paid leave. The difference in the annual hours measure therefore lets usual weekly hours drive the variation in the difference between the two measures. Alternatively, both measures could be adjusted to account for weeks worked rather than weeks paid. Because I am interested in the difference between the two measures, adjusting both down by the same amount will not influence the final measure of work off the clock. All statistics showing annual hours levels reflect the lack of adjustment.

1.3.1 Sample Construction

For the final merged ACS-LEHD analysis sample, I first separately prepare ACS respondents and create an analogous person-year-quarter frame for the LEHD. The ACS preparation begins by attaching a protected identification key (PIK) to each ACS respondent. A PIK is a person-level identifier that allows one to link ACS responses to other individual datasets within the U.S. Census Bureau.[23] Using an internal crosswalk, I link ACS responses from 2010 to 2013. For each year, an ACS respondent links to a PIK at rates between 91% and 94% per year. After deduplicating records within a year, I am left with 18.3 million ACS responses.

I merge the person-year-quarter LEHD frame and the ACS to create the final analysis dataset. The resulting sample contains 571,000 records. The small sample is a result of

[23]See Wagner and Layne (2014) for a description of the U.S. Census Bureau's PIK assignment process.

limiting the sample of ACS respondents to those who have positive hours paid in the previous year in the LEHD hours-reporting states from the time of their ACS interview. I further restrict the sample to ensure that the LEHD and the ACS frames accord as close as possible. For consistency with the LEHD, I restrict ACS respondents to those age sixteen and over, and I require that they have no jobs in other states over the previous year by consulting the standard LEHD EHF, which includes all available states.[24] Using the ACS-reported dominant job, I exclude federal government employees.[25] I use the ACS reported residence and exclude all respondents who neither live in an hours reporting state, nor in a border state. It is perfectly reasonable for an ACS respondent living in North Dakota, for example, to work in Minnesota. As a result of these restrictions, I reduce the sample to 438,000 observations.

The final sample contains additional restrictions to negate any anticipated frame differences, which could lead to biased measurement of work off the clock. I restrict the sample to records who report working a full year (50-52 weeks) in the ACS, and who report positive hours in the LEHD for every quarter in the past year. This restriction is reasonable for a few reasons. First, the ACS sample restriction narrows weeks worked considerably. The weeks worked variable in ACS is binned, and the bins become coarser the fewer weeks one works.[26] The LEHD restriction to working in the reference quarter as well as the preceding four quarters simply ensures consistency with the ACS. Finally, I drop observations where usual weekly hours is imputed, and observations where workers receive more than 20 percent of their income from self-employment earnings as reported on the ACS. The final dataset contains 218,000 records.

[24] I also exclude respondents for whom I find jobs with zero or missing hours data. This is evidence of unit non-response and would bias measures of work off the clock.

[25] The ACS-reported dominant job does not conform to the definition of a dominant job in the LEHD. For the ACS, the dominant job is the main job in the week prior to the ACS response. Given the stability of federal jobs in general, and the high tenure of my final sample, I use the two definitions interchangeably. Checks for consistency find a high degree of agreement.

[26] There is research pointing to quality problems in hours worked for the ACS for part-time workers (Baum-Snow and Neal, 2009). I include full-year part-time workers, though all results are robust to their exclusion with some loss of precision.

In order for the final analysis sample to remain representative of the United States population, I adjust the ACS sample weights. When I merge the ACS to the LEHD universe of job records, the ACS weights are no longer representative of the U.S. population due to differences in frame. I first use inverse probability weighting to adjust the ACS sample weights for PIKs missing at random in the sense of Rubin (1987).[27] Second, I adjust the sample weights to match national demographic characteristics in the 2009-2013 ACS for full-year workers excluding federal employees. I adjust based on age, gender, race/ethnicity, and education. The resulting sampling weights allow for inferences about the population after linking the ACS to a different universe.

1.3.2 Variable Construction

I create the final total for hours paid from the LEHD over the previous year by summing hours over jobs and weighting the interview quarter and ending quarter by the ACS interview date. For person i employed at job j at any year-quarter t between 2010 and 2013, define $h_{j(i),t}$ as the gross quarterly hours for person i in job j in quarter t. I consider only LEHD jobs between the quarter of the ACS interview (t), and four quarters prior ($t-4$), inclusive. Total hours paid over the previous year therefore include five quarters of data, which is one too many. To calculate the final annual hours paid in the LEHD over the previous year from the survey interview date, I sum hours over all jobs, and take a weighted average of hours in the interview quarter and the last quarter,

$$H_{i,lehd} = (1-\rho)\left(\sum_{i\in J} h_{j(i),t-4}\right) + \left(\sum_{t=1}^{3}\sum_{i\in J} h_{j(i),t-k}\right) + \rho\left(\sum_{i\in J} h_{j(i),t}\right). \tag{1.1}$$

The first term on the right-hand side of equation 1.1 is the sum of all hours paid to respondent i across all jobs J in which the respondent worked in period $t-4$. This sum

[27] This is also known as missing conditional on observable covariates. See appendix A.2 for details on inverse probability weighting.

is multiplied by the weight $1 - \rho$. The middle term is the sum of hours paid at all jobs in the three quarters immediately preceding the interview quarter. The last term is the sum of hours across all jobs in the interview quarter multiplied by weight ρ. The weights are based on the percentage of the interview quarter in scope for the total hours calculation.[28]

To construct annual hours worked in the ACS, I assume that usual weekly hours is equivalent to average weekly hours and then multiply usual weekly hours by 50, 51 and 52 weeks to get three measures of annual hours worked. Table 1.1 provides statistics for annual hours worked assuming a 52 week work year. This is my preferred hours measure for several reasons. First, the final sample contains workers with relatively high tenure (over six years), and I perform checks for continuous employment in the LEHD over the previous four quarters from the reference quarter. Second, the ACS asks for usual weeks worked including paid sick days, paid vacation, and military service. Given checks for continuous employment and because the ACS weeks worked question includes paid leave, 52 weeks seems the most reasonable measure of weeks paid that accords with the LEHD.

The final dependent variable of interest is the log difference between annual hours worked from the ACS, and total hours paid over the previous year from the LEHD. Recall that the sample is restricted to full-year workers, and that ACS weeks worked is binned for full-year workers. Denote the annual ACS hours measure $H^w_{i,acs}$ where $w \in \{50, 51, 52\}$ is the possible weeks worked. The annual LEHD hours measure is denoted $H_{i,lehd}$. The measure of difference between hours worked and hours paid is $y^w_i = \ln(H^w_{i,acs}) - \ln(H_{i,lehd})$. I construct the log difference for all three ACS measures of ACS annual hours. I then winsorize at the 5% and 95% level in order to mitigate bias induced by extreme outliers.[29]

[28]For example, if an ACS respondent completed the survey on May 10th, the weight assigned to the interview quarter would be equal to the 40 days in the quarter divided by 91, which is total days in the second quarter. The weight on the end month is simply one minus the interview quarter weight.

[29]The following analysis has also been carried out with a dependent variable winsorized at the 1% and 99%. Results are qualitatively unchanged.

Table 1.1: Summary Statistics for Analysis Sample

	mean	sd
ACS annual hours (52 Weeks)	2,079	500.3
LEHD annual hours	1,946	504.8
Annual hours error (50 weeks, ACS)	0.030	0.251
Annual hours error (51 weeks, ACS)	0.055	0.241
Annual hours error (52 weeks, ACS)	0.079	0.237
Firm/Job Characteristics		
Unemployment rate (%)	7.354	1.901
Private, for-profit firm	0.743	0.437
Supervisory, Non-production	0.275	0.447
Top Quartile, Liklihood Not Paid by Hour	0.277	0.448
Dominant job tenure (quarters)	26.96	21.0
Demographic Characteristics		
Age	42.30	12.87
Male	0.519	0.500
Non-white	0.239	0.426
Bachelor's degree or higher	0.325	0.468

Notes: $N = 218,000$, with 58 commuting zones. Annual hours error is the difference between log hours worked in the ACS and log hours paid from the LEHD. The ACS hours paid measure is defined by multiplying the usual weekly hours by the number of weeks paid. Supervisory workers adhere to the Bureau of Labor Statistics definition of supervisory or non-production workers. See text for details.

Demographic, firm, and job characteristics come from a combination of the ACS and the LEHD. Many characteristics are available in both the LEHD and the ACS. I use the ACS for demographic characteristics, which are occasionally imputed in the LEHD. For the residence, I use the ACS reported residence. The LEHD residence is from a fixed time period each year, and will not necessarily correspond to the residence at the time of ACS interview. I use firm characteristics from the LEHD dominant job[30] as it comes from an

[30]I define the LEHD dominant job as the job with the most hours paid in the three quarters which lie

administrative source and best hues towards my preferred definition of a dominant job. I make use of whether or not a worker is paid by the hour. This is not available in either the ACS or LEHD.[31] I use the Current Population Survey to impute hourly/non-hourly pay using industry, occupation, and earnings according to the ACS. I then bin the resulting probability of non-hourly pay into quartiles for use in the empirical analysis.[32]

1.3.3 Summary Statistics

Summary statistics of the mean and standard errors for the final sample are found in Table 1.1. The first two lines give the summary statistics for estimates of annual hours worked from the ACS and annual hours paid from the LEHD. The bottom panel displays demographic characteristics used to match the analysis sample to the U.S. population. The restriction to full-year workers is perhaps the most salient for what follows. Note that the average tenure in LEHD dominant jobs is a little under 27 quarters, or slightly greater than 6.5 years.

Figure 1.1 shows the full distribution of the difference in log hours worked from log hours paid. I use my preferred ACS hours worked measure, which assumes 52 paid weeks. The distribution is centered slightly to the right of zero, but it is highly skewed with a long right tail implying many more people report working more than their employers say they do. Note that the distribution is winsorized at the 95% level, which slightly truncates the right tail of the distribution, but analysis relaxing the winsorization to the 99% level does not alter the results.

completely within the preceding year from the date of the ACS interview.

[31] Deciphering salaried workers in the LEHD is not as simple as looking at low variance in quarterly hours across quarters for a given job. Due to the abundance of weekly or bi-weekly pay periods, jobs with constant weekly hours paid will nonetheless display quarter to quarter variance in hours paid as the number of pay periods in a quarter fluctuates.

[32] See appendix A.3 for details.

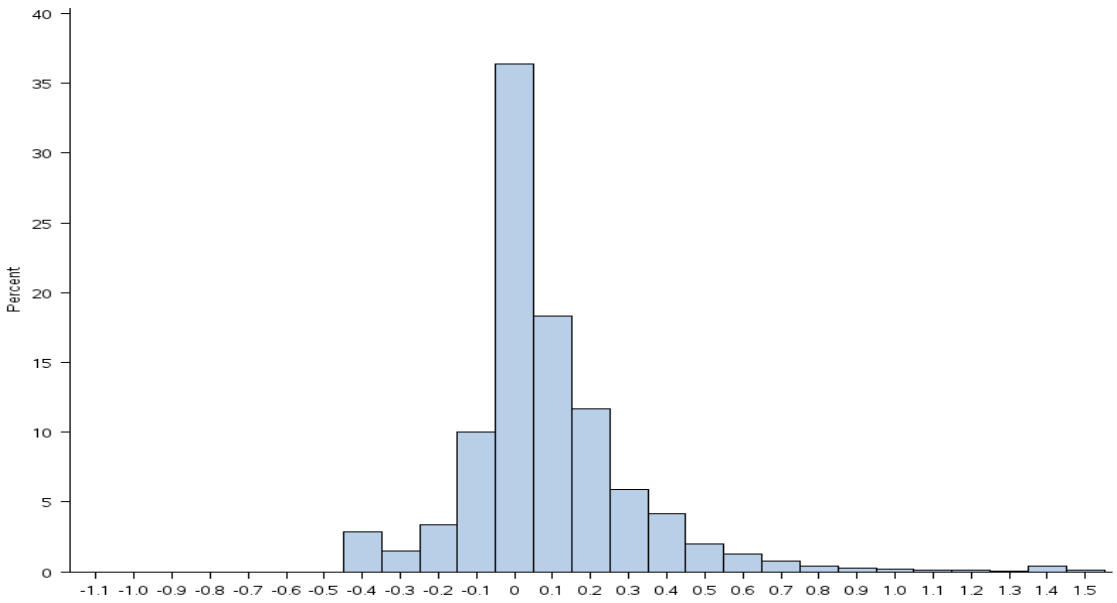


Figure 1.1: Distribution of the Difference of Log Hours Worked and Log Hours Paid

Notes: Variable is the difference in log ACS hours worked from log LEHD hours paid for the analysis sample, winsorized at the 5% and 95% level. $N = 218,000$. See Table 1.1 for summary statistics.

Partitioning the distribution by a few characteristics reveals large differences between hours worked and hours paid, particularly for those who report working more than 40 hours per week. Figure 1.2 shows the large disparity in work off the clock by workers who self-report working more than 40 hours per week. The figure partitions the distribution of the difference in log hours into those who self-report usually working less than 40 hours per week (top panel), those who usually work 40 hours per week (middle panel), and those who usually work more than 40 hours per week (bottom panel). The difference is stark. Those who work less than 40 hours per week show a small difference in hours worked compared to hours paid compared with those who usually work exactly 40 hours per week, with means of 0.031 [0.268] and 0.032 [0.179], respectively. Most of the mass is centered around zero in both distributions, with a slight right skew. In contrast, for those who report working more than 40 hours the distribution shifts to the right with the mass less sharply concentrated. The mean for this distribution is 0.21 [0.238].[33]

[33]For comparison, Figure A.1 in appendix A.4 shows the earnings error for the same two groups. Here,

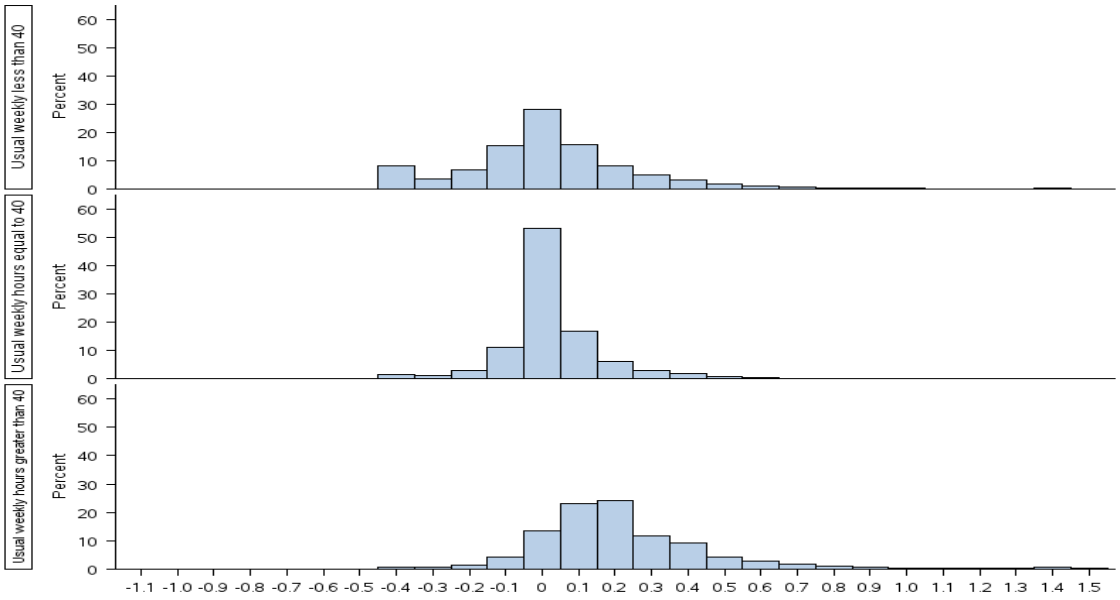


Figure 1.2: Distribution of the Difference of Log Hours Worked and Log Hours Paid by ACS usual weekly hours

Notes: Variable is the difference in log ACS hours worked from log LEHD hours paid for the full sample, winsorized at the 5% and 95% level. The variable is partitioned by whether an ACS respondent answers that she usually works either 1) less than 40 hours per week (top panel) or 2) exactly 40 hours per week (middle panel) or 3) more than 40 hours per week (bottom panel). Top panel $N = 49,000$, middle panel $N = 108,000$, bottom panel $N = 61,000$. Mean of top panel 0.031 [0.268], middle panel 0.032 [0.179] and bottom panel 0.21 [0.238].

Figure 1.3 displays the distribution of log hours worked less log hours paid by LEHD hours paid. I divide the annual hours paid measure from the LEHD by 52 weeks to obtain average weekly hours paid. Figure 1.3 partitions the log difference distribution into those who on average were paid less than 40 hours per week (top panel), 40 hours per week (middle panel), and more than 40 hours per week (bottom panel). The top two panels show a significant right skew in the distribution – what would be predicted from Figure 1.2. However, firms who report paying for over 40 hours per week on average have a more symmetric distribution with a mean of -0.020 [1.18]. In general, Figure 1.3 reinforces the finding that hours worked and hours paid accord quite closely, but that hours worked over 40 hours per week are not well recorded by employers.

the distributions show a significant left skew, but in general they are quite similar.

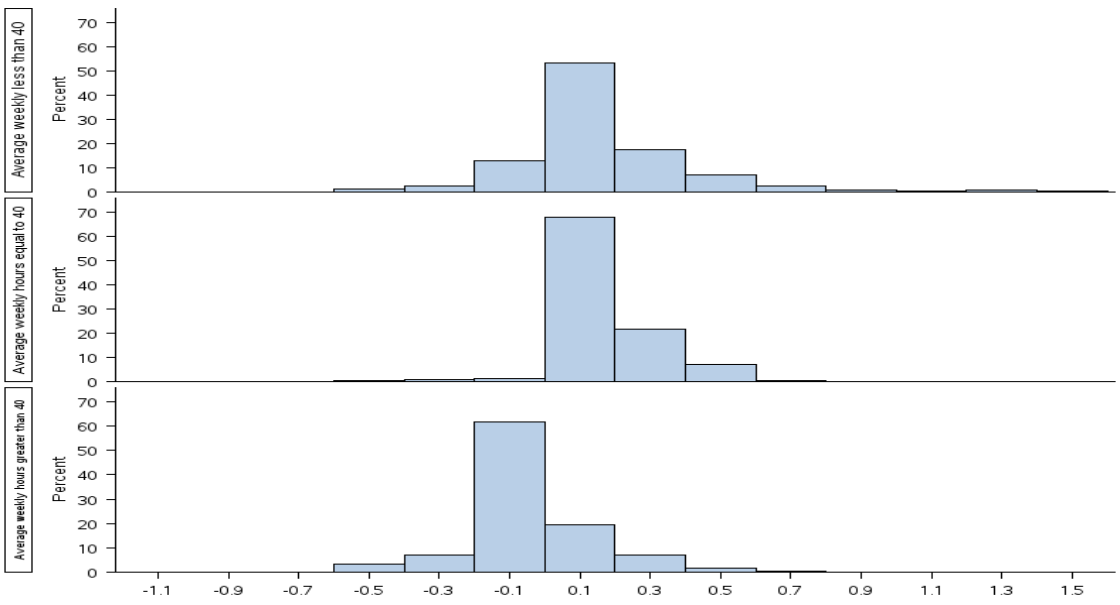


Figure 1.3: Distribution of the Difference of Log Hours Worked and Log Hours Paid by LEHD Hours Paid

Notes: Variable is the difference in log ACS hours worked from log LEHD hours paid for the full sample, winsorized at the 5% and 95% level. The variable is partitioned by average weekly hours paid (annual LEHD hours divided by 52). Top panel: less than 40 hours per week. Middle panel: exactly 40 hours a week. Bottom panel: more than 40 hours per week. Top panel $N = 116,000$, middle panel $N = 24,000$, bottom panel $N = 79,000$. Mean of bottom panel 0.158 [1.91], middle panel 0.108 [1.09] and bottom panel −0.020 [1.18].

The distributions of log hours worked less log hours paid suggest that hours worked is accurately reported except for workers who report working more than 40 hours per week. Recent studies support this finding by comparing survey responses from the Current Population Survey (CPS) to the American Time Use Survey (ATUS). The ATUS is a time use survey, and it is generally thought to more accurately reflect hours worked due to its short recall period. Frazis and Stewart (2009) compare hours worked per job in the two surveys at the person level. They find that survey responses from the CPS of hours worked are remarkably close to the ATUS, with any overstatement confined to multiple jobholders.[34] The conclusion of these validation studies is that respondents report work hours accurately in surveys.

[34]See Frazis and Stewart (2010) and Frazis and Stewart (2004) for similar results.

1.4 Hours & the Business Cycle

1.4.1 Empirical Strategy

Differentiating between the hypothesis of extra effort from employees in the form of off the clock work when labor markets are slack versus the competing labor hoarding hypothesis is the key empirical question. The ideal design for such an analysis would randomly assign different unemployment rates, or other measures of labor market slack, to many identical self-contained economies and then observe the co-movement of hours paid and hours worked. Such a fantasy experiment is not feasible. This paper uses variation in local labor market slack to infer the relationship between the business cycle and the difference between hours paid and hours worked. Specifically, I use the variation in the change in unemployment rates in commuting zones to identify hours off the clock. This approach takes advantage of the large geographic dispersion of the United States, which effectively partitions the country into many self-contained regional economies.[35]

I use the ACS respondent's place of residence for the local labor market defined as a commuting zone. Given the relatively small geographic area of four states and their adjoining neighbors, I use commuting zones as the primary local labor market unit as it classifies all counties into a commuting zone. The metropolitan statistical area (MSA) is also an appealing geographic delineation for a local labor market as it is defined as a collection of counties around a major (or minor) city usually including its suburbs. However, the MSA excludes some mostly-rural counties from any MSA, which leads to further reductions in sample size.[36]

I use the unemployment rate to measure the local labor market from the Local Area

[35]Schaller (2016) is a recent example who employs a similar approach.
[36]Headline results are robust to to local labor markets defined by MSA.

Unemployment Statistics (LAUS) from the Bureau of Labor Statistics (BLS). LAUS provides local area unemployment rates derived from BLS surveys and unemployment insurance data. The data are available monthly for small areas including MSAs and counties. I use county unemployment rates to construct monthly commuting zone unemployment rates by averaging county unemployment rates for each county in a commuting zone weighting the average by the labor force of each county. I then average monthly commuting zone unemployment rates into a quarterly commuting zone unemployment rate. I use the quarterly commuting zone unemployment rate corresponding to the quarter of the ACS survey response as an indicator of labor market conditions. The LAUS unemployment rates are known to be noisy. By averaging over the months in the reference quarter, I allow the data to accord to the underlying analysis sample, and eliminate some of the noise.

Figure 1.4 shows the time series of quarterly unemployment rates for the two largest commuting zones by population in the analysis sample. One commuting zone contains Minneapolis, Minnesota and the other contains Seattle, Washington. The unemployment rates for both commuting zones rise quickly at the onset of the Great Recession before gradually declining. Seattle's commuting zone increases to almost 10% at its peak before declining to below 7% at the end of the analysis sample. Minneapolis's unemployment rate peaks at over 7% before 2010, and then declines to slightly below 5% at the end of 2013. For comparison, the national unemployment rate declined from 9.8% in January 2010 to 6.6% in January 2014. All results that follow should be interpreted with these general macroeconomic conditions in mind. Although this is just two commuting zones, Figure 1.4 shows that there is ample variation both within and across commuting zones in the unemployment rate.[37]

[37]Figure A.3 shows the full distribution of unemployment rate changes across all commuting zones in the analysis sample.

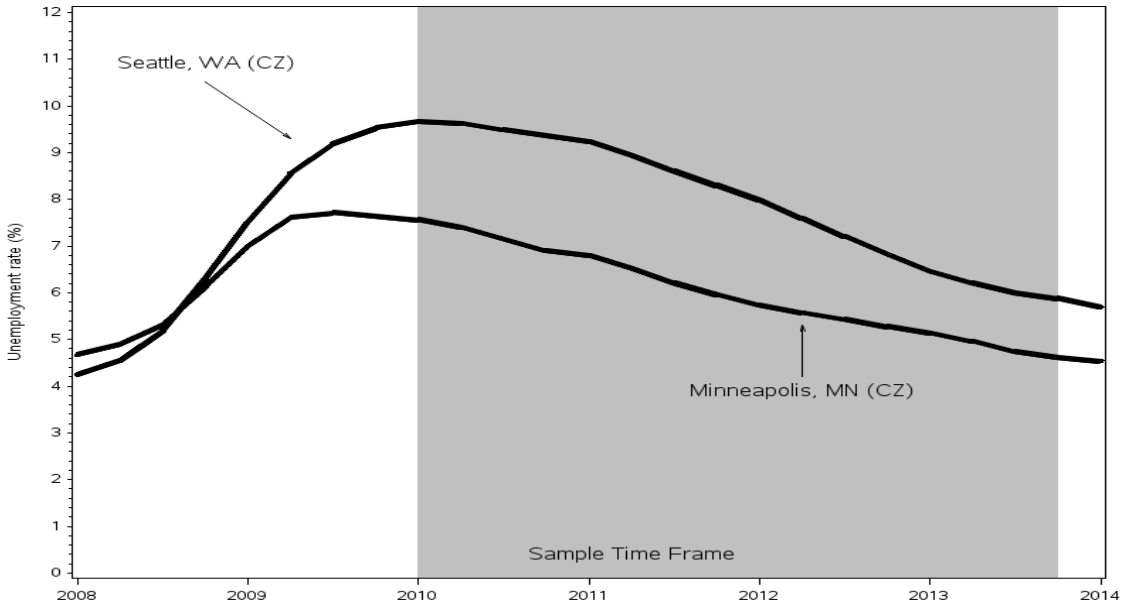


Figure 1.4: Quarterly unemployment rates, Seattle, WA and Minneapolis, MN Commuting Zones

Notes: Seattle, WA (CZ) refers to the entire commuting zone which contains the city of Seattle, Washington. Minneapolis, MN refers to the entire commuting zone which contains the city of Minneapolis, Minnesota. The shaded region marks the time frame of the analysis sample.

The specification for the ordinary least squares (OLS) estimate is given by,

$$y_{i,t}^{52} = \beta U_{cz,t} + \delta \mathbf{X}_i + \psi \mathbf{J}_{j(i)} + \alpha_s + \omega_t + \phi_{cz} + \epsilon_{i,t,cz} \quad . \tag{1.2}$$

where $y_{i,t}^{52}$ is the difference between the logarithm of ACS annual hours worked and the logarithm of annual LEHD hours paid. I use my preferred 52-week measure for ACS hours worked in the dependent variable. For this linear specification, the choice of dependent variable corresponding to different weeks worked will not matter for the final estimates. Differences in the dependent variable due to varying weeks worked only shift the intercept of the regression line and have no effect on the slope, and therefore the coefficient of interest.

The variable $U_{cz,t}$ is the unemployment rate in the commuting zone of the residence of person i in interview quarter t. The vector $\mathbf{X}_i$ captures demographic characteristics of person i, while $\mathbf{J}_{j(i)}$ captures job and firm characteristics of person i employed at dominant

job j. Job characteristics include tenure, industry fixed effects and an indicator variable for whether a worker is employed in a supervisory or non-production occupation.[38]

I also include fixed commuting zone effects. Denoted ϕ_{cz}, the inclusion of fixed commuting zone effects identifies the effect of the unemployment rate on the difference of log hours using time-series variation within commuting zones. The inclusion of commuting zone time trends in some specifications identifies the effect using de-trended time series variation within commuting zones. I include additional controls for fixed year-quarter and state fixed effects in the specification denoted ω_t and α_s, respectively. Finally, $\epsilon_{i,t,cz}$ is the error term.

Although commonly used in the literature, the use of local unemployment rates presents some problems for measures of labor market slack. The unemployment rate confounds both supply and demand induced responses of labor force participation. Changes in the unemployment rate may be endogenous to other variables forcing changes in work off the clock. In this setting, where the dependent variable is the difference in log hours worked from log hours paid, such endogenous changes are harder to envision, but it is not implausible that changes in state or local labor programs to encourage labor force participation may also change an employer's unemployment insurance hours reporting requirements.[39]

In order to buttress the results using the local unemployment rate, I also construct an employment shift-share measure of plausibly exogenous labor demand. This shift-share index commonly credited to Bartik (1991), but used extensively in local labor market analyses,[40] uses a local labor market's industrial composition in a base year to predict employment growth in the local labor market in subsequent years. The intuition behind the in-

[38] I use the BLS definition for production and non-supervisory workers, which is defined by industry and occupation. See U.S. Bureau of Labor Statistics (2004) for a detailed description.

[39] It appears Rhode Island began requiring employers to report hours paid at the same time it began a new labor market policy. Whether the former is in response to the latter has proven difficult to pin down.

[40] See Blanchard and Katz (1991) Bound and Holzer (2000) Autor and Duggan (2003) for other examples.

strument is to fix local industrial composition, and allow national employment growth to predict local employment growth. If drivers of national growth are applied uniformly, local labor markets with greater concentrations of the growth industries should see greater predicted employment growth simply due to their industrial composition. I follow Autor and Duggan (2003) and construct predicted employment growth in labor market cz at time t from base year t_0 as

$$\hat{G}_{t,cz} = \sum_k \delta_{t_0,cz,k} G_{t,k} \quad .$$

The first term, $\delta_{t_0,cz,k}$, gives the share of employment in NAICS sector k in local labor market cz at time t_0, and the second term, $G_{t,k}$, is the change in log national employment in NAICS sector k between the base year and time t. I exclude the local labor market cz for the computation of national growth rates for each local labor market.[41]

1.4.2 Results

The results of the specification employed in equation 1.2 are presented in Table 1.2. All standard errors in parentheses are cluster-robust, clustered by commuting zone (Cameron and Miller, 2015). Columns (1) to (4) add time trends, firm/job characteristics, and demographic characteristics to the regressions, respectively. Column (4) shows my preferred specification including commuting zone specific time trends, as well as time varying firm, job and demographic controls. The preferred specification has a coefficient (β) of -0.00191 (0.000767), which is small but precisely measured.

The estimated sign of the coefficient suggests labor hoarding best explains the data. As labor markets become tighter (the unemployment rate decreases), hours worked expand faster than hours paid (the wedge between hours worked and hours paid increases). To

[41] National employment growth rates are from the Bureau of Labor Statistics CEW, and I construct local labor market shares from the U.S. Census Bureau's QWI, which is benchmarked to the CEW. I use 2007 for the base year because it is the peak of previous business cycle, though results are robust to using year 2000.

Table 1.2: The Effect of the Unemployment Rate on Work Off the Clock

	(1)	(2)	(3)	(4)
Unemployment rate (β)	-0.00143*	-0.00198**	-0.00182**	-0.00191**
	(0.00075)	(0.00081)	(0.00075)	(0.00077)
State FE	X	X	X	X
Commuting Zone Time Trends		X	X	X
Firm & Job controls			X	X
Demographic controls				X
R^2	0.006	0.007	0.093	0.104

Notes: $N = 218,000$, with 58 commuting zones. Dependent variable is the difference between log annual ACS hours calculated at 52 weeks and log annual LEHD hours. All regressions estimated using ordinary least squares. Cluster-robust standard errors clustered by commuting zone. Stars on standard errors accord to p-values as follows: *** p<0.01, ** p<0.05, * p<0.1.

provide a valid explanation for the counter-cyclical change in productivity, the sign on the coefficient would be positive. The magnitude of the coefficient indicates a small cyclical component to off the clock work. In my preferred specification, a one point decline in the unemployment rate results in an increase in off the clock work of 0.1%. The OLS results find a relationship that lacks a strong cyclical component, and therefore cannot explain the cyclical change in productivity.

The instrumental variables estimates confirm the OLS results. These results are presented in Table 1.3 with my preferred specification contained in panel A, column (3). Panel B shows the first stage results. The Bartik shift-share instrument is highly correlated with the unemployment rate across all specifications. The coefficient on predicted employment growth is −27.8 (3.99), with the sign in the correct direction (higher predicted employment growth leads to a lower unemployment rate), and a first stage F-statistic of 48.43. The two stage least squares coefficient on the unemployment rate in column (3) is still negative and larger in absolute value than in the OLS specification in column (2). The coefficient on unemployment is now −0.00274 (0.00154). The instrumental variables

Table 1.3: The Effect of the Unemployment Rate on Work Off the Clock, Two Stage Least Squares Results

	(1)	(2)	(3)
	Panel A: 2SLS Results		
Unemployment rate (β)	-0.00286	-0.00254	-0.00274*
	(0.00177)	(0.00159)	(0.00154)
	Panel B: First Stage Results		
Bartik Shift-Share	-29.07***	-27.78***	-27.80***
	(3.790)	(3.987)	(3.994)
First stage F-Statistic	58.84	48.57	48.43
State FE	X	X	X
Commuting Zone Time Trends		X	X
Firm & Job controls			X
Demographic controls			X
R^2	0.005	0.006	0.104

Notes: $N = 218,000$, with 58 commuting zones. Dependent variable is the difference between log annual ACS hours calculated at 52 weeks and log annual LEHD hours. Panel A shows two stage least squares estimates. Panel B shows the corresponding first stage regressions. Cluster-robust standard errors clustered by commuting zone. Stars on standard errors accord to p-values as follows: *** p<0.01, ** p<0.05, * p<0.1.

estimates come with the cost of a loss in precision with the coefficient only significant at the 10% level. The instrumental variables results provide evidence for supply-induced responses possibly attenuating the OLS results.

The headline results of a limited cyclical component mask heterogeneity in off the clock work and the business cycle. Table 1.4 fits the same model given in equation 1.2, but interacts the unemployment rate with subgroups most likely to work off the clock.[42] Column (1) shows the slopes of the regression lines associated with production and non-supervisory workers and their complement, supervisory workers. Supervisory workers are most likely to be paid a fixed salary and seem likely candidates to be driving off the

[42]Table A.5 shows qualitatively similar results by running the regression separately for each group.

clock work.[43] Somewhat surprisingly, column (1) shows that it is production and non-supervisory workers driving the results. The coefficient estimate is −0.00219 (0.00075). In contrast, the estimate for the coefficient for supervisory workers is imprecise and slightly positive, 0.00111 (0.00080).

Table 1.4: The Effect of the Unemployment Rate on Work Off the Clock, Heterogeneous Effects

	(1)	(2)	(3)
$U_{cz,t}$ * Non-supervisory	-0.00219***		
	(0.00075)		
$U_{cz,t}$ * Supervisory	-0.00108		
	(0.00108)		
$U_{cz,t}$ * Less than Bachelor's Degree		-0.00214***	
		(0.00085)	
$U_{cz,t}$ * Bachelor's Degree or higher		-0.00164*	
		(0.00095)	
$U_{cz,t}$ * Least likely non-hourly pay			-0.00228**
			(0.00079)
$U_{cz,t}$ * Most likely non-hourly pay			-0.00091
			(0.00099)
p-value (from F-test coefficients are equal)	0.17	0.59	0.09
R^2	0.104	0.105	0.103

Notes: $N = 218,000$, with 58 commuting zones. Dependent variable is the difference between log annual ACS hours calculated at 52 weeks and log annual LEHD hours. Non-supervisory is an indicator for observations who meet the definition of a production or non-supervisory worker according to the occupation and industry of her dominant job. Supervisory refers to the compliment of non-supervisory. The definition of non-supervisory or production worker comes from the Bureau of Labor Statistics. "Least likely non-hourly pay" refers to observations who are below the median in likelihood they are not paid by the hour. "Most likely non-hourly pay" refers to observations above the median likelihood not paid by the hour. All regressions run using OLS with the same specification as Table 1.2 column (4). Cluster-robust standard errors clustered by commuting zone. Stars on standard errors accord to p-values as follows: *** p<0.01, ** p<0.05, * p<0.1.

Similar results obtain when interacting the unemployment rate by method of pay and by skill level. Column (3) in Table 1.4 interacts the results by whether the ACS respondent is likely paid by the hour. Workers most unlikely to be paid hourly have a coefficient on

[43]This is due to the duties test for exemption from overtime in the FLSA. A key component of the test is whether an employee works in a supervisory capacity.

the slope of the regression line of 0.00049 (0.00091). In contrast, workers most likely be paid by the hour have an estimated coefficient of −0.00214 (0.000852). Column (2) shows the results for workers with and without a bachelor's degree. Although there is no *a priori* reason why workers with a bachelor's degree would be more or less likely to work off the clock, in practice a bachelor's degree is highly correlated with supervisory work and non-hourly pay arrangements. The point estimate for workers without a bachelors degree is qualitatively similar to the coefficient on non-supervisory workers, and precise, obtaining an estimate of −0.00220 (0.00079). The results confirm that it is lower skilled workers likely paid by the hour who are driving the results.

The preceding results show support for labor hoarding driving the cyclical component of off the clock work, though the effect is small. Another further confirmation of the labor hoarding hypothesis is the tenure of workers. Firms holding onto excess labor in a downturn will be eager to deploy it once demand picks up. In contrast, in an efficiency wage setting it seems plausible that workers with strong attachment to particular jobs who also happen to retain their dominant job after the Great Recession are particularly good matches with their employers, and their high tenure precludes them from feeling threatened with layoffs.[44]

Table 1.5 shows the results interacting various firm and job characteristics with the local unemployment rate. Column (1) gives the results for tenure, where the estimating equation augments equation 1.2 with indicator variables for length of tenure one year or less, 1-3 years, 3-5 years, and greater than 5 years. The estimated slopes show that it is longer tenured workers driving the results with point estimates of −0.00230 (0.00786), −0.00223 (0.00803), and −0.00170 (0.00754), for workers with 1-3 years, 3-5 years, and greater than 5 years of tenure, respectively. These coefficient estimates are all significant, and much larger in magnitude than for workers who have less than one year of tenure.

[44]This is one result of the model of Rebitzer (1987).

Table 1.5: Regression Results for Unemployment Rate and Work Off the Clock, Heterogeneous Effects, Additional Results

	(1)	(2)	(3)
$U_{cz,t}$ * Tenure: 1 year or less	-0.00058		
	(0.00096)		
$U_{cz,t}$ * Tenure: 1-3	-0.00230***		
	(0.00078)		
$U_{cz,t}$ * Tenure: 3-5	-0.00223***		
	(0.00080)		
$U_{cz,t}$ * Tenure: +5	-0.00170**		
	(0.00075)		
$U_{cz,t}$ * Firm size: 0-19		0.00052	
		(0.00083)	
$U_{cz,t}$ * Firm size: 20-49		-0.00203**	
		(0.00090)	
$U_{cz,t}$ * Firm size: 50-249		-0.00241***	
		(0.00083)	
$U_{cz,t}$ * Firm size: 250-999		-0.00310***	
		(0.00086)	
$U_{cz,t}$ * Firm size: 1,000-2,499		-0.00147*	
		(0.00077)	
$U_{cz,t}$ * Firm size: + 2,500		-0.00238***	
		(0.00079)	
$U_{cz,t}$ * Firm age: 0-1			0.00105
			(0.00124)
$U_{cz,t}$ * Firm age: 2-3			0.00000
			(0.00119)
$U_{cz,t}$ * Firm age: 4-5			0.00000
			(0.00105)
$U_{cz,t}$ * Firm age: 6-10			-0.00060
			(0.00103)
$U_{cz,t}$ * Firm age: +11			-0.00211***
			(0.00075)
R^2	0.104	0.105	0.104

Notes: $N = 218,000$, with 58 commuting zones. Dependent variable is the difference between log annual ACS hours calculated at 52 weeks and log annual LEHD hours. Tenure is measured in years on the dominant job. All regressions run using OLS with the same specification as Table 1.2 column (4). Cluster-robust standard errors clustered by commuting zone. Stars on standard errors accord to p-values as follows: *** p<0.01, ** p<0.05, * p<0.1.

Table 1.5 also shows that off the clock work is concentrated in large, old firms. The results suggest a labor hoarding model where firms work the excess labor they have kept during a downturn harder during the ensuing recovery.[45] The results further suggest that although likely paid a salary, higher skill workers likely have better outside options and are better able to resist pressures to vary hours according to cyclical labor market pressure.

1.4.3 Robustness Checks

The first robustness check tests for sensitivity of the results to the specification of the dependent variable. The first two columns of Table 1.6 run regression 1.2 using alternate specifications of work off the clock. Column (1) shows the results using the disparity between hours worked and hours paid according to Tornqvist et al. (1985).[46] This measure is defined if either hours measure is equal to zero, and is roughly equivalent to the log measure of the percent difference. I also do not winsorize this variable. Column (2) uses a binary indicator variable for the dependent variable, which equals unity if hours worked in the ACS exceeds hours paid from the LEHD. This makes equation 1.2 a linear probability model. The point estimates on the unemployment rate for columns (1) and (2) are −0.00183 (0.00080) and −0.00505 (0.00204), respectively. The estimates are precise, and indicate that the results are not sensitive to the specification of the dependent variable.

[45] Table A.6 shows the correlation between off the clock work and firm employment growth. There is little correlation. This is not inconsistent with a model of firms with large fixed adjustment costs who work their existing workforce as long and hard as possible before eventually adjusting.

[46] Formally, this is $y_i^{alt} = \frac{H_{i,acs} - H_{i,lehd}}{\frac{1}{2}(H_{i,acs} + H_{i,lehd})}$. Within the economics literature, this measure is usually credited to Haltiwanger et al. (1996).

Table 1.6: The Effect of the Unemployment Rate and Work Off the Clock, Robustness Checks

	(1) DHS	(2) Binary	(3) Interview Quarter	(4) Log Diff.	(5) Log Diff.	(6) Log Diff.
Unemployment rate (β)	-0.00183**	-0.00505**	0.0182***			
	(0.00080)	(0.00204)	(0.0023)			
Year/Year job growth				0.0024***		
				(0.0004)		
$U_{cz,t}$ * Full-time					-0.00148*	
					(0.00084)	
$U_{cz,t}$ * Part-time					-0.00192	
					(0.00125)	
$U_{cz,t}$ * More than one job						-0.00152
						(0.00155)
$U_{cz,t}$ * One job						-0.00172**
						(0.00070)
p-value (from F-test coefficients are equal)					0.71	0.88
R^2	0.085	0.057	0.110	0.074	0.104	0.117

Notes: $N = 218,000$, with 58 commuting zones. Dependent variable in column (1) is Haltiwanger et al. (1996) measure, column (2) is an indicator for whether ACS annual hours exceeds LEHD annual hours, and column (3) is the difference in log ACS hours measured in the ACS interview quarter, and log LEHD hours in the ACS interview quarter. For columns (4)-(6) dependent variable is the difference between log annual ACS hours calculated at 52 weeks and log annual LEHD hours. Year/Year Job Growth is defined for the commuting zone using the Haltiwanger et al. (1996) measure calculated from the Quarterly Workforce Indicators. Full-time is defined as usual weekly hours greater than or equal to 35 in the ACS. Part-time is less than 35 usual weekly hours in the ACS. More than one job is defined as holding more than one job over the previous year from the ACS interview in the LEHD. All regressions run using OLS with the same specification as Table 1.2 column (4). Cluster-robust standard errors clustered by commuting zone. Stars on standard errors accord to p-values as follows: *** p<0.01, ** p<0.05, * p<0.1.

The next threat to identification comes from recall bias. The ACS asks the respondent for usual weekly hours over the previous year. My measure of work off the clock assumes "usual" weekly hours is equivalent to average weekly hours, to respondents.[47] Hours are growing in the analysis sample as the unemployment rate trends down between 2010 and 2013. It is possible that workers simply report their usual hours right around the time of interview, and not over the previous year. If this is the case, then my results will show a positive relationship between work off the clock and tightening of the local labor market – exactly what I find. To test for this, in column (3) of Table 1.6 I use the difference between log hours worked and log hours paid in the interview quarter.[48] If ACS respondents understand usual hours to mean hours over the previous year, this new measure should yield a positive coefficient – the average over the previous year will be smaller than the interview quarter. If respondents are giving usual hours in the interview quarter, the point estimate on the unemployment rate should be close to zero and/or imprecisely measured. The positive estimate in column (3) of 0.0182 (0.0023), suggests that respondents interpret usual weekly hours as asking for average hours in the analysis time frame.

Table 1.6, column (4) tests whether the results are sensitive to the measure of labor market slack. The Local Area Unemployment Statistics from the BLS are model based, and known to be noisy. In column (3) I use the year-over-year job growth rate for the commuting zone calculated using the approximate log change credited to Haltiwanger et al. (1996). The year over year growth rate is measured from one year prior to the interview quarter. The data come from the Quarterly Workforce Indicators (QWI) from the U.S. Census Bureau. The QWI are the public-use version of the LEHD, and are therefore derived from administrative unemployment insurance data. The coefficient estimate us-

[47]The exact question is (capitalization as in original), "During the PAST 12 MONTHS, in the WEEKS WORKED, how many hours did this person usually work each WEEK?"

[48]I use the log of gross hours in the interview quarter for hours paid and the log of usual weekly hours scaled by the number of weeks in the quarter for hours worked.

ing the year over year job growth rate is 0.0024 (0.0041). The estimate is precise and the sign of the coefficient is in the correct direction. That is, a positive coefficient for local employment growth is consistent with a negative coefficient for the local unemployment rate.[49]

Another potential problem with my specification is that some jobs may not be reported to the unemployment insurance system at all. If the ACS respondent includes the hours from these jobs in usual weekly hours, this could bias the results to look like labor hoarding.[50] The construction of the analysis sample makes this scenario unlikely, but I test for off the books work by interacting the unemployment rate by full time status and multiple job holding.[51] The assumption is that workers who hold multiple jobs and/or work part-time are more likely to pick up short, informal jobs that are not reported to the unemployment insurance system. The results are given in columns (5) and (6) of Table 1.6. The coefficients for the slope of part-time workers and multiple job holders are 0.00044 (0.001199) and −0.00155 (0.00154), respectively. Neither estimate is precise, though the coefficient on multiple job holders suggests that it is not implausible off the books work may be influencing the results.

[49]Results are robust to sensitivity of included commuting zones, and the later inclusion of Oregon in 2012, though with a loss of precision in the latter case. Due to disclosure risks, these results are not able to be released at this time.

[50]From the perspective of a statistical agency calculating productivity, this distinction between "off the books" and "off the clock" is likely irrelevant.

[51]I calculate multiple job holding summing the number of jobs in the LEHD over the previous year.

1.5 Hours & Labor Compliance

1.5.1 Empirical Strategy

Somewhat surprisingly, this paper finds that off the clock work has a small pro-cyclical component, which is driven by low-skill workers likely paid by the hour. Non-compliance with labor market regulations offers a possible explanation for these results. Non-compliance can either be explicit, by paying workers under the table, or refusing to pay over-time for hours worked over 40 hours per week. There are also subtle ways this can arise. For example firms may mis-classify employees who should be non-exempt as exempt, and shift more hours to these workers. To test this theory, we need to use the characteristics of firms. In particular, theory and survey evidence suggest that small firms are more likely to engage in non-compliance behavior due to lower costs of bankruptcy, diminished reputation, and lower productivity firms.[52]

The relevant facts presented in Section 1.2 indicate that off-the-clock work, and wage and hour violations often times operate through the (mis)management of hours of work. Even in the case of minimum wage violations, non-hourly pay frequency and its associated ambiguity of work hours is correlated with FLSA violations. In this section I test for whether hours-based evidence for work off the clock is present in my representative microdata. I use ordinary least squares regression and follow the specification of Bernhardt et al. (2013) using survey data, as well as Ji and Weil (2015) who use administrative data.

Before analyzing firm characteristics in greater depth, I study the following specification in order to ensure that key variables behave roughly as expected. The model is,

$$y_{i,t}^{52} = \delta \mathbf{X}_i + \psi \mathbf{J}_{j(i)} + \alpha_s + \omega_t + \epsilon_{i,t,cz} \quad , \tag{1.3}$$

[52] Milkman et al. (2012) and Mendeloff et al. (2006) provide recent examples.

where the vector $\mathbf{X}_i$ captures demographic characteristics of person i, while $\mathbf{J}_{j(i)}$ captures job and firm characteristics of person i employed at dominant job j. The specification is close to equation 1.2, except for the omission of commuting zone effects. The empirical strategy used to identify off the clock work for labor compliance does not rely on variation in local labor market conditions. I therefore drop the commuting zone effects as well as commuting zone time trends.

In addition to the standard firm and individual controls, I augment the model with indicators for firm size. Both Bernhardt et al. (2013) and Ji and Weil (2015) emphasize the role played by firm size in labor violations. I therefore augment equation 1.3 vector $\mathbf{J}_{j(i)}$ with bins for firm size given by,

$$\sum_{b=1}^{B} \psi_b \mathbf{1}\{k_b \leq firmsize_{j(i)} < k_{b+1}\} \quad ,$$

where b indexes the bins with B total bins, and k is the set of bounds defining the bins with $B+1$ bounds. The indicator function takes a value of one if $firmsize_{j(i)}$ – the firm size of the LEHD dominant job – falls within the specified firm size category. The firm-person match which constitutes a job in the LEHD uses a definition of a firm as a state-level tax reporting entity. It is not uncommon for a larger national entity to be the real owner of a firm, with the state distinction a product of the state-based nature of unemployment insurance records. The LEHD remedies this by bringing in firm size from the U.S. Census Bureau's Longitudinal Business Database (LBD). All firm size categories use the LBD's definition of a firm taking into account inter-state ownership.

1.5.2 Results: Firm Size

The estimation results for equation 1.3 are given in Table 1.7. My preferred specification given in column (1) lines up well with prior research. Workers with a bachelors degree,

men, and workers whose main job is with a private, for profit firm are more likely to report more hours worked than hours paid. Two curious results are that the model indicates that people of color report fewer hours worked than hours paid, and U.S. citizens slightly more hours worked than hours paid. The higher incidence of work off the clock for U.S. citizens is likely due to the fact that previous survey evidence included workers who are not able to legally work in the U.S. The analysis sample includes only workers who are found in the administrative data. Inclusion in the UI data generally necessitates a social security number suggesting that non-citizens in the sample are likely different than non-citizens in purely survey data. For workers of color, the difference in sign has no easy explanation, other than previous results found only a tenuous relationship between race and labor compliance.

The coefficients of greatest interest are on the indicator variables for supervisory workers and on the quartiles of likelihood a worker is not paid by the hour. It is generally assumed that employers pay little attention to the hours for supervisory workers because they are exempt from overtime and usually not paid by the hour. In such cases one should expect hours worked to exceed hours paid. The coefficient on the indicator for supervisory workers in the main specification in column (1) is 0.00909 (0.0050) indicating that supervisory workers, all else equal, work 0.9% more hours than those for which they are paid over the course of the year. At first glance this seems low, but it is important to realize that this indicator, based on BLS definitions of supervisory workers, is highly correlated with my imputation of non-hourly pay probability.

I measure non-hourly pay probability and its association with work off the clock by indicators for quartiles of probability of non-hourly pay. In Table 1.7 column (1) the indicators for quartile of non-hourly pay are measured relative to the lowest quartile, where workers most are likely to be paid by the hour. The results indicate that workers most likely to be in non-hourly pay arrangements are the most likely to report more hours

Table 1.7: Characteristics of Work Off the Clock: OLS Results

	(1)	(2)
Private, for-profit firm	0.0125***	0.0135***
	(0.00281)	(0.00497)
Not White	-0.0270***	-0.0483***
	(0.00563)	(0.00306)
Hispanic	-0.00869*	-0.00969
	(0.00522)	(0.00624)
Male	0.00445**	0.00830***
	(0.00180)	(0.00142)
U.S. Citizen	0.0319***	0.0203***
	(0.00584)	(0.00269)
Bachelors degree of higher	0.0486***	0.0572***
	(0.00175)	(0.00316)
Supervisory Worker	0.00909*	-0.00970*
	(0.00502)	(0.00566)
Second Quartile	-0.0463***	
	(0.00346)	
Third Quartile	-0.0355***	
	(0.00538)	
Top Quartile	0.0136**	
	(0.00591)	
Firm Controls	X	X
Year-Quarter FE	X	X
State FE	X	X
Demographic Controls	X	X
Observations	218,000	67,000
R^2	0.073	0.096

Notes: Dependent variable is the difference between log annual ACS hours calculated at 52 weeks and log annual LEHD hours. Quartile is the quartile of likelihood paid non-hourly. Supervisory worker is defined by observation's industry and occupation according to BLS definition of production and non-supervisory workers. Column (2) subsets the regression to only observations in the top quartile of likely non-hourly pay. Cluster-robust standard errors clustered by state-firm. Stars on standard errors accord to p-values as follows: *** p<0.01, ** p<0.05, * p<0.1.

worked than hours paid. The coefficient in this case is 0.0136 (0.0059). Column (2) in the same table fits equation 1.3 using OLS but subsets the data to only include those in the top quartile of likelihood of non-hourly pay. Results are in general qualitatively unchanged, though the coefficient on the indicator for supervisory flips sign – it is now negative – indicating the limited explanatory power of this distinction in comparison to the pay type, although these two effects are difficult to distinguish.

Table 1.8: Summary Statistics for Analysis Samples by Firm Size

Firm Employment Size	**0-19**		**20-49**		**50-249**		**250-999**		**1,000-2,499**		**+2,500**	
	mean	sd	mean	sd	mean	sd	mean	sd	mean	sd	mean	sd
ACS annual hours (52 Weeks)	1,973	584.4	2,054	536.6	2,090	486.6	2,096	462.4	2,117	452.6	2,104	483.5
LEHD annual hours	1,800	572.3	1,912	533.2	1,965	494.5	1,990	493.6	1,963	480.8	1,977	475.2
Annual hours error (50 weeks, ACS)	0.055	0.284	0.035	0.256	0.025	0.246	0.017	0.240	0.043	0.253	0.024	0.243
Annual hours error (51 weeks, ACS)	0.084	0.271	0.062	0.244	0.050	0.238	0.042	0.232	0.068	0.241	0.049	0.231
Annual hours error (52 weeks, ACS)	0.109	0.270	0.086	0.242	0.074	0.237	0.066	0.229	0.091	0.238	0.073	0.225
Firm/Job Characteristics												
Unemployment rate (CZ)	0.074	2.008	7.292	1.955	7.331	1.955	7.326	1.928	7.423	1.854	7.327	1.817
Private, for-profit firm	0.856	0.352	0.812	0.391	0.745	0.436	0.613	0.487	0.668	0.471	0.754	0.431
Likely Exempt occupation (Management)	0.283	0.450	0.287	0.452	0.273	0.446	0.265	0.441	0.248	0.432	0.280	0.449
Top Quartile, Liklihood Not Paid by Hour	0.176	0.381	0.258	0.438	0.260	0.439	0.265	0.441	0.320	0.466	0.322	0.467
Dominant Job tenure (quarters)	23.84	19.43	26.14	20.76	26.44	20.43	28.09	20.95	28.13	21.21	27.81	21.81
Demographic Characteristics												
Age	41.79	13.73	41.43	13.48	42.26	12.99	43.11	12.59	43.15	12.62	42.20	12.46
Male	0.528	0.499	0.563	0.496	0.532	0.499	0.507	0.500	0.494	0.500	0.509	0.500
Non-white	0.163	0.369	0.181	0.385	0.213	0.409	0.250	0.433	0.270	0.444	0.280	0.449
Bachelors degree or higher	0.226	0.418	0.266	0.442	0.279	0.449	0.318	0.466	0.352	0.478	0.393	0.488
Observations	24,000		18,000		35,000		31,000		17,000		75,000	

Notes: $N = 218,000$. Annual hours error is the difference between log hours worked in the ACS and log hours paid from the LEHD. The ACS hours paid measure is defined by multiplying the usual weekly hours by the number of weeks paid in each row. Firm size employment groups are based on employment firm size on the 12th of the month of the year of ACS interview.

Table 1.8 presents the summary statistics for the analysis sample by firm size. First, ACS and LEHD average annual hours generally increase as firm size increases, though the mean difference between log hours worked and log hours paid decline slightly as firm size increases. Characteristics of firms and workers employed show more variation by firm size. Slightly more than 85% of firms in the smallest firm size category (0-19) work at private, for profit firms compared to slightly more than 75% in the largest firm size category (firms with more than 2,500 employees). In addition, workers in the smallest firms are much more likely to be paid by the hour. Firms with 0-19 employees employ 17.6% of workers in the analysis sample in top quartile of non-hourly pay probability compared to 32.0% and 32.2% in the largest two firm size categories, respectively. Finally, 22.6% of workers in the smallest firms have at least a bachelors degree compared to 39.3% in the largest firm size category.

The results of including firm size in equation 1.3 are given in Table 1.9. The coefficients correspond to the indicator variables of their respective firm sizes. All coefficients should be interpreted as the increased (decreased) difference in hours worked compared to hours paid in firms with $2,499$ employees or more. The results in column (2) show that workers whose primary job is in small firms appear to work more hours than they are paid. The coefficient on the 0-19 employment category is positive at 0.0231 (0.0053) and precise. The same results are displayed graphically in Figure 1.5.

One potential issue with this specification is the worry that firm size bins are arbitrarily chosen. In addition, I have continued to use the winsorized difference in logs as the measure of off the clock work. I check for the robustness of both of these concerns in additional specifications. The first two columns of Table 1.10 use finer firm size bins and reach largely similar conclusions. The results from column (2) are also presented graphically in Figure 1.6. The coefficients on all three firm size bins with employment less than 20 employees are significant and positive, and of similar magnitude to the coefficients in Table

Table 1.9: Off the Clock Work by Firm Size, Coarse Firm Size Bins

	(1)	(2)
0-19	0.0358***	0.0231***
	(0.0057)	(0.0053)
20-49	0.0132**	-0.0002
	(0.0063)	(0.0097)
50-249	0.00117	0.0009
	(0.0057)	(0.0046)
250-999	-0.0070	-0.0075
	(0.0063)	(0.0050)
1,000-2,499	0.0180**	0.0032
	(0.0083)	(0.0066)
Firm controls		X
Year-Quarter FE		X
State FE		X
Demographic controls		X
R^2	0.003	0.116

Notes: $N = 218,000$. Dependent variable is the difference between log annual ACS hours calculated at 52 weeks and log annual LEHD hours. All coefficients reported in reference to largest firm size group; firms with employment greater than or equal to 2,500. Cluster-robust standard errors clustered by state-firm. Stars on standard errors accord to p-values as follows: *** p<0.01, ** p<0.05, * p<0.1.

1.9. The exception is the coefficient on the smallest firm size bin, 1-4, is a little less than twice the next two firm size bins. Coefficients in this specification should be interpreted again in reference to the largest firm size, which is now firms with employment greater than 10,000. In addition to the finer firm size categories, columns (3) and (4) in Table 1.10 show results for the same specification but using Haltiwanger et al. (1996) measure of differences in hours worked compared to hours paid. Even though this measure is not winsorized, it shows quantitatively similar results to previous specifications.

The finding that smaller firms are associated with higher self-reported hours worked

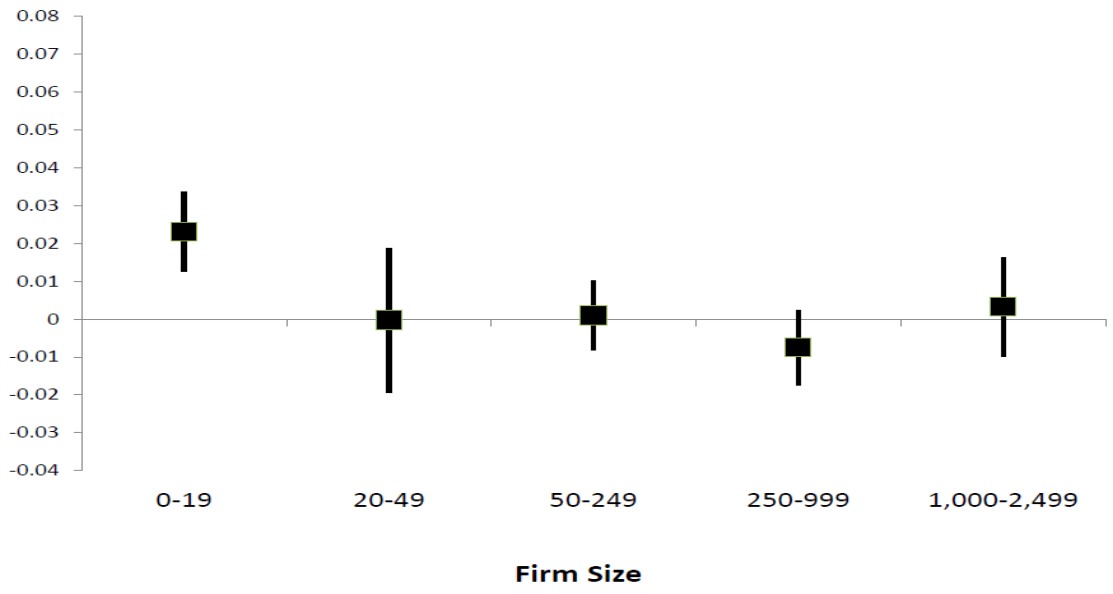


Figure 1.5: Regression Coefficients for LEHD Firm Size Categories

Notes: Boxes are regression coefficient point estimates and lines are 95% confidence intervals. All results are relative to the largest firm size category, +2,500 employment. See Table 1.9 column (3).

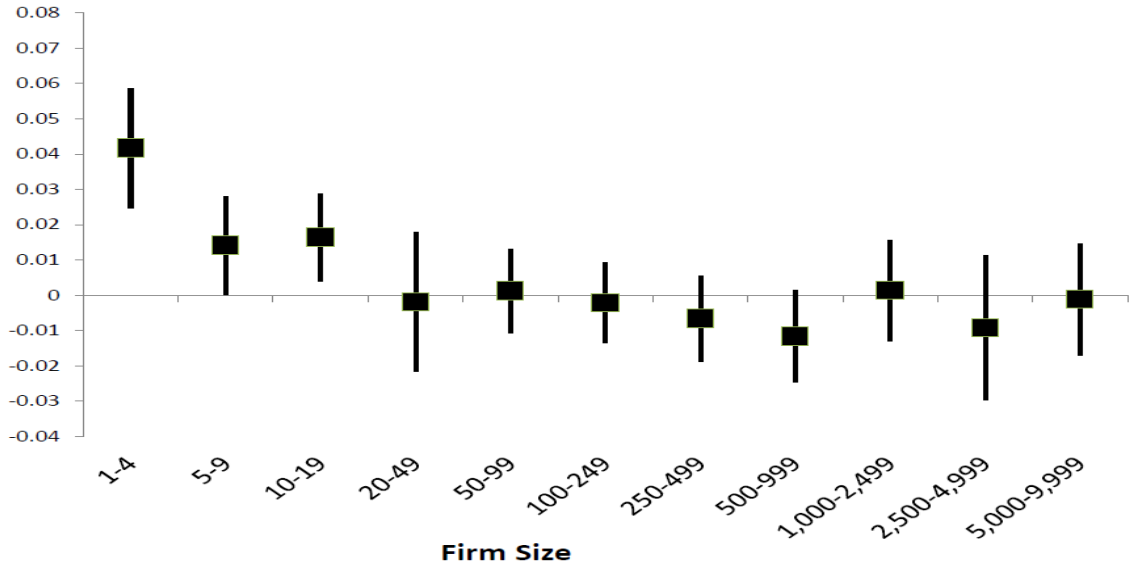


Figure 1.6: Regression Coefficients for BDS Firm Size Categories

Notes: Boxes are regression coefficient point estimates and lines are 95% confidence intervals. All results are relative to the largest firm size category, +10,000 employment. See Table 1.10 column (2).

compared to hours paid is consistent with the current literature on non-compliance, but masks some of the story. Recall that the previous results center around low-skill workers, who are more likely to be paid an hourly wage. To test whether the firm size results are being driven by supervisory (likely exempt) or non-supervisory workers, I run the same specification with the firm size indicators, but now I interact the firm size indicator variables with an indicator for supervisory employees.

I plot the results graphically in Figures 1.7 and 1.8. The results are also available in Table A.2 column (1). All results are shown relative to non-supervisory workers in the largest firm size category. What the two Figures make clear is that the greater off the clock work in the smallest firm size category appears to be driven by workers in non-supervisory jobs. Figure 1.7 plots the estimated coefficients and their associated 95% confidence intervals for all firm sizes interacted with supervisory workers. The key point from this Figure is the uniformity and statistical significance of almost all coefficients. The estimated coefficients range from 0.0198 to 0.599. Given the near uniformity across firm size categories, it seems unlikely that supervisory workers are driving the results by firm size.

In contrast, the estimated coefficients on non-supervisory workers depict a different pattern. Figure 1.8 plots these coefficients, which are also available in Table A.2 column (1), top panel. The estimated coefficients are uniformly small, and an estimate of zero cannot be rejected at the 5% confidence level. The estimate for the smallest firm size category is the only positive point estimate, 0.0294 (0.0294). The pattern on the coefficients for the firm size indicators follows the same pattern as firm size overall. The results provide support for the predictions of fewer hours reporting in smaller firms, and wage and hour compliance. The effect is not driven by supervisory or non-production workers, rather by production or non-supervisory workers.

Figures 1.9 and 1.10 provide further evidence for non-supervisory workers explaining

Table 1.10: Off the Clock Work by Firm Size, Fine Firm Size Bins

	(1)	(2)	(3)	(4)
1-4	0.0586***	0.0418***	0.0505***	0.0425***
	(0.0096)	(0.0086)	(0.0106)	(0.0095)
5-9	0.0302***	0.0142**	0.0295***	0.0205**
	(0.0080)	(0.0070)	(0.0093)	(0.0081)
10-19	0.0241***	0.0164***	0.0250***	0.0228***
	(0.0078)	(0.0063)	(0.0092)	(0.0074)
20-49	0.0119	-0.0018	0.0125	0.00247
	(0.0081)	(0.0101)	(0.0095)	(0.0110)
50-99	0.00165	0.0012	0.0052	0.0071
	(0.0079)	(0.0060)	(0.0091)	(0.0070)
100-249	-0.0013	-0.0021	0.0000	0.0012
	(0.0078)	(0.0057)	(0.0091)	(0.0067)
250-499	-0.0057	-0.0065	-0.0029	-0.00278
	(0.0082)	(0.0061)	(0.0096)	(0.0071)
500-999	-0.0104	-0.0116*	-0.0081	-0.0088
	(0.0089)	(0.0066)	(0.0101)	(0.0075)
1,000-2,499	0.0167*	0.0013	0.0182*	0.0033
	(0.0097)	(0.0072)	(0.0109)	(0.0080)
2,500-4,999	-0.0036	-0.0091	0.0000	-0.0052
	(0.0118)	(0.0105)	(0.0136)	(0.0117)
5,000-9,999	-0.0042	-0.00117	-0.0006	0.0013
	(0.0103)	(0.0081)	(0.0113)	(0.0086)
Firm controls		X		X
Year-Quarter FE		X		X
State FE		X		X
Demographic controls		X		X
R^2	0.004	0.116	0.002	0.099

Notes: $N = 218,000$. Dependent variable for columns (1) and (2) is the difference between log annual ACS hours calculated at 52 weeks and log annual LEHD hours. For columns (3) and (4) dependent variable is Haltiwanger et al. (1996) change. All coefficients reported in reference to largest firm size group; firms with employment greater than 10,000. Cluster-robust standard errors clustered by state-firm. Stars on standard errors accord to p-values as follows: *** p<0.01, ** p<0.05, * p<0.1

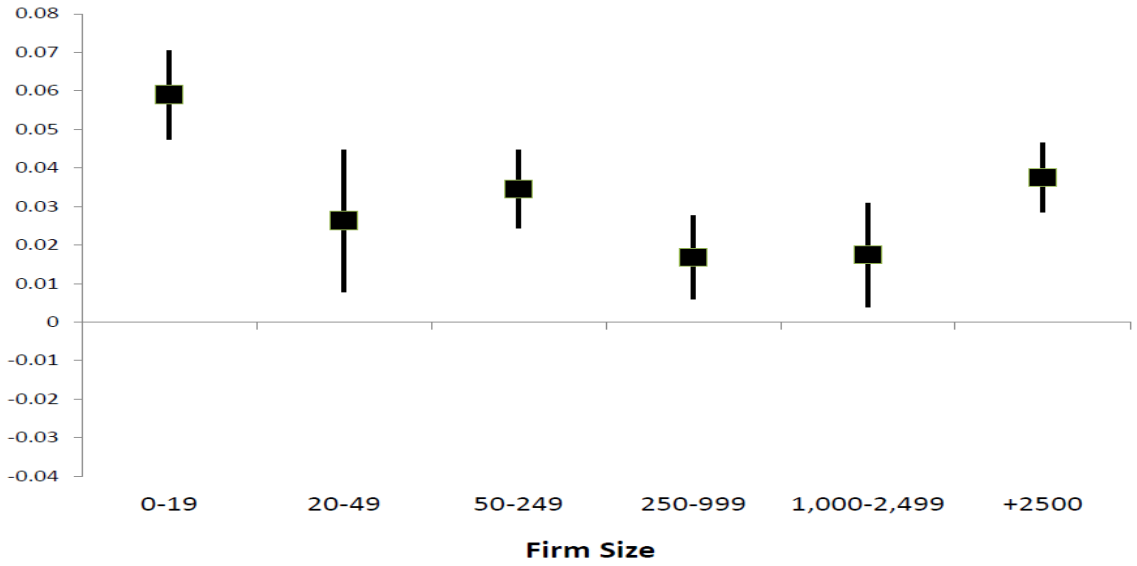


Figure 1.7: Regression Coefficients for Firm Size: Supervisory Workers

Notes: Boxes are regression coefficient point estimates and lines are 95% confidence intervals. All results are relative to the largest firm size category, +2,500 employment, for production and non-supervisory workers. See Table A.2 column (1), bottom panel.

the firm size effect. Figure 1.9 shows the results of firm size interacted with an indicator for the bottom half of hourly pay distribution. The coefficients are similar in magnitude to the coefficients interacting firm size with and indicator for non-supervisory workers. Turning to Figure 1.10, the case of the top half of the probability of hourly pay distribution, the results are again similar to non-supervisory workers in Figure 1.8. This is somewhat curious as we would expect non-hourly workers to be driving the variation. Due to the imputation of pay status, this cannot be explicitly ruled out. It is possible that those who are unlikely to be paid hourly are driving the variation. What is clear is that the results indicate that small firms overwhelming report more hours worked than hours paid, and that the results are driven by low-skill production and non-supervisory workers.

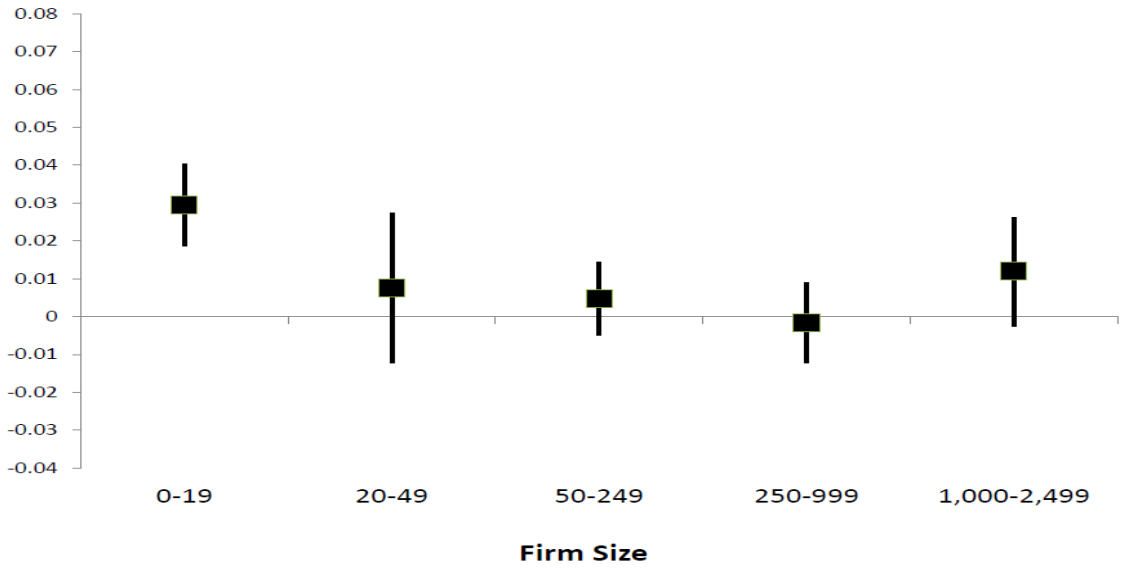


Figure 1.8: Regression Coefficients for Firm Size: Non-supervisory and Production Workers

Notes: Boxes are regression coefficient point estimates and lines are 95% confidence intervals. All results are relative to the largest firm size category, +2,500 employment, for production and non-supervisory workers. See Table A.2 column (1), top panel.

1.5.3 Results: Industry

In addition to firm size, industry is often a strong predictor of hours off the clock. More generally, one implication of true failure to record and track employee hours should be wage and hour violations. Table 1.11 displays the top ten NAICS three-digit industries ranked by their share of all Department of Labor investigative actions where a violation was found to have occurred between 2010 and 2013. The top ten industries account for 62.6% of all violations, and comprise 36.1% of private employment, on average, between 2010 and 2013. Violations are further concentrated in the top three industries: Food Services & Drinking Places, Specialty Trade Contractors (roofers, for example), and Administrative & Support Services (temporary employment services, janitors, and security guards). In general, low-wage service industries account for the majority of wage and hour violations.

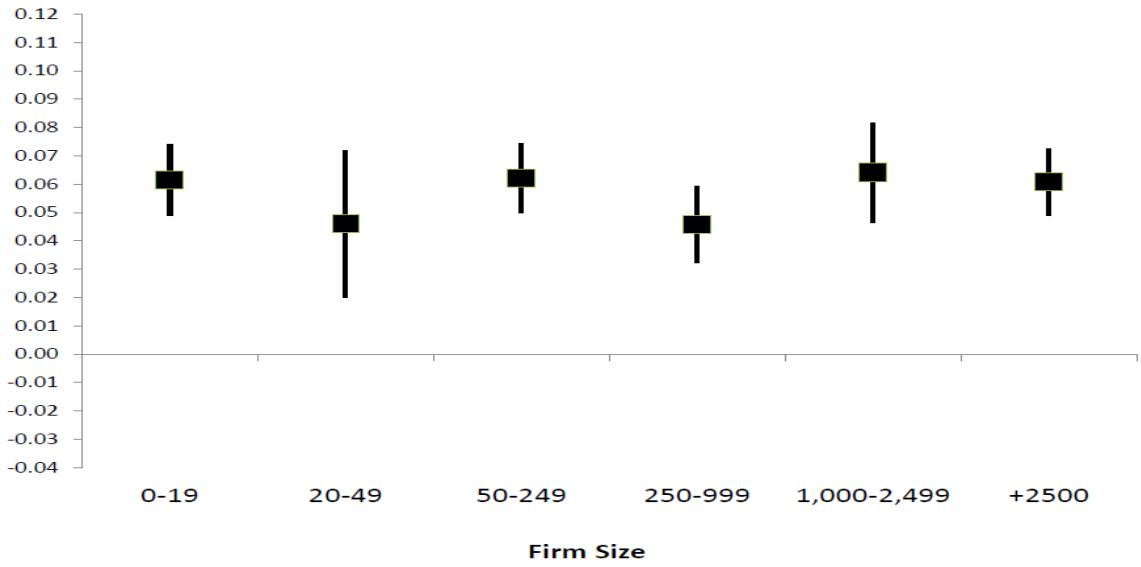


Figure 1.9: Regression Coefficients for Firm Size: Worker's Likely Paid a Salary, Top 50%

Notes: Boxes are regression coefficient point estimates and lines are 95% confidence intervals. All results are relative to the largest firm size category, +2,500 employment, for production and non-supervisory workers. See Table A.2 column (2), bottom panel.

If the difference between hours worked and hours paid is indicative of employers working low-wage workers more than they report, this should be concentrated in the industries displayed in Table 1.11. To test this hypothesis, I create three indicator variables that evaluate to unity if an ACS respondent's LEHD dominant job is in one of the top ten, top five, or top three NAICS 3-digit industries by share of enforcement actions, respectively. I fit equation 1.3 separately for each of the three indicator variables including demographic, job, firm, and local labor market controls. The hypothesis is that the progression to industries with higher concentrations of wage and hour violations should yield increased incidence of off the clock work.

The results of this exercise are displayed in Table 1.12. The first column shows the results of the indicator variable, denoted "High Incidence of Violation", for the top ten industries. The coefficient estimate is 0.0016 (0.0037), indicating workers report working 0.16% more hours in these industries, although with little precision. Moving to column (2)

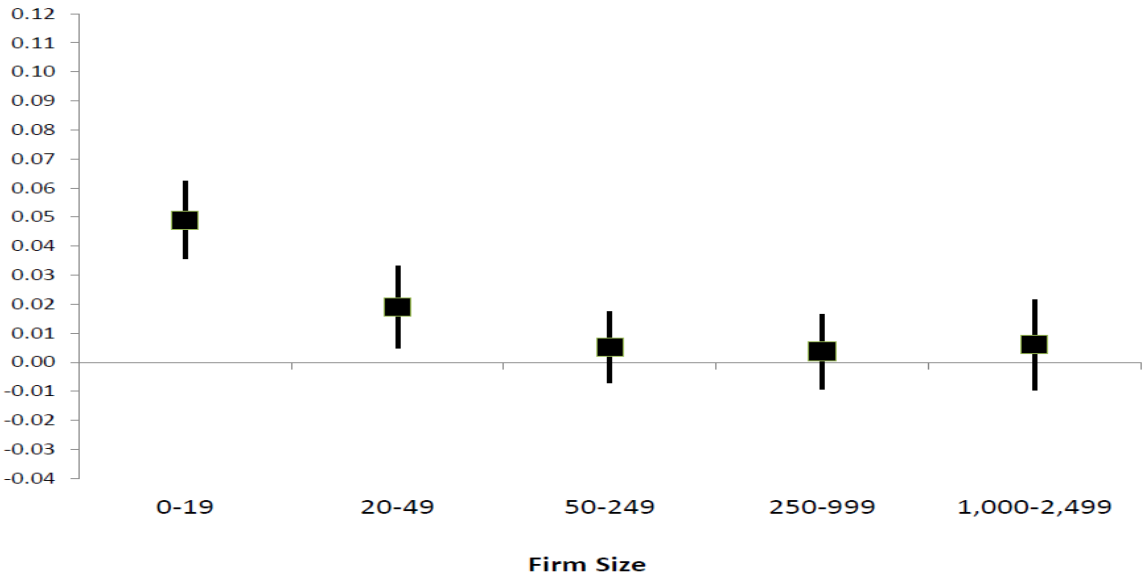


Figure 1.10: Regression Coefficients for Firm Size: Worker's Likely Paid a Salary, Bottom 50%

Notes: Boxes are regression coefficient point estimates and lines are 95% confidence intervals. All results are relative to the largest firm size category, +2,500 employment, for production and non-supervisory workers. See Table A.2 column (2), top panel.

and concentrating on the top 5 industries, the estimate is now precise and 1.54% (0.0037) higher than in other industries. Finally, column (3) focuses on the top 3 industries, which account for over a third of enforcement actions. Again, the coefficient estimate is precise and increases, as expected, with workers in these industries reporting working 3.04% (0.0035) more than they are paid.

1.6 Conclusion

This study uses a unique dataset of person-level survey responses of hours worked linked to the person's employer reports of hours paid. To the best of my knowledge, this is the first study in United States linking hours worked to administrative reports of hours paid at the person-level. I interpret this as a measure of work off the clock. I use the measure

Table 1.11: Industries with Largest Share of Wage and Hour Violations, 2010-2013

Industry (NAICS, 3-digit)	Share of Wage and Hour Violations (%)		Overall share (%), QCEW	
	Actions	Employees	Establishments	Employment
Food Services & Drinking Places (722)	22.8	22.6	6.5	8.9
Specialty Trade Contractors (238)	6.0	5.8	5.5	3.2
Administrative & Support Services (561)	5.9	10.2	5.1	6.8
Social Assistance (624)	5.2	3.6	3.6	2.4
Accommodation (721)	5.0	3.7	0.7	1.6
Nursing & Residential Care Facilities (623)	4.8	5.0	0.8	2.9
Ambulatory Health Care Services (621)	3.9	3.2	6.3	5.7
Gasoline Stations (447)	3.3	1.4	1.2	0.8
Food and Beverage Stores (445)	3.2	2.0	1.6	2.6
Construction of Buildings (236)	2.6	1.9	2.6	1.1
Total				
Cumulative share of total (%)	62.6	59.4	33.9	36.1
Cumulative count	34,100	398,000		
Mean count (2010-2013)			2,980,000	39,400,000

Notes: Wage and Hour violations from U.S. Department of Labor, Wage & Hour division. Sample is all compliance actions with findings beginning and ending between 2010 and 2013, and where at least one wage and hour violation occurred. Employees is share of all affected employees in actions with findings of at least one violation. Shares of establishments and employment are the average shares between 2010-2013. Data are from the Bureau of Labor Statistics, Quarterly Census of Employment and Wages of all private establishments.

Table 1.12: Off the Clock Work by Industry

	(1) Top 10	(2) Top 5	(3) Top 3
High Incidence of Violation	0.0016	0.0154***	0.0304***
	(0.0037)	(0.0037)	(0.0035)
R^2	0.137	0.137	0.138

Notes: $N = 218,000$. Dependent variable is the difference between log annual ACS hours calculated at 52 weeks and log annual LEHD hours. "High Incidence of Violation" is an indicator variable, which evaluates to unity if the LEHD dominant job is in the top 10, 5, or 3 NAICS 3-digit industries ranked by incidence of wage and hour violations. See Table 1.11 for ranking and details. All regressions estimated using OLS, and include firm, job, demographic, and local labor market controls. Cluster-robust standard errors clustered by state-firm. Stars on standard errors accord to p-values as follows: *** p<0.01, ** p<0.05, * p<0.1.

to test various theories of employer-employee bargaining.

I differentiate efficiency wage and labor hoarding theories of hours bargaining by testing the association of off the clock work with cross-section variation in local labor market slack. The hypothesis that greater work time off the clock may have a significant impact on measured productivity is not supported by the empirical findings of this study. The results find support for labor hoarding, though the effect is small. I find further evidence that the cyclical component of work off the clock is driven by low-skill workers likely paid by the hour.

The results also contribute to the emerging literature on labor compliance, and specifically on more explicit determinants of work off the clock. In line with previous studies, I find evidence that smaller firms are more likely to have employees report higher hours worked compared to hours paid. The results indicate that for the smallest firms, work off the clock is driven by production and non-supervisory employees.

At a broader level this study argues that hours of work should become a more prominent topic in labor economic research. The measurement of hours is important for the study of wages and productivity, and the often times casual tracking of hours by firms and workers leads to failures of labor market compliance. Finally, this paper also advises caution when using data on hours. The greater emphasis the economics profession is placing on administrative data combined with casual hours reporting by firms may produce misleading results when using administrative data on hours. More stringently tracking hours worked for both hourly and non-hourly workers would be an appropriate policy response. This will help both administrative data collection and analysis, and may also have the beneficial effect of greater wage and hour compliance.

CHAPTER 2

HOURS ADJUSTMENTS: EVIDENCE FROM LINKED EMPLOYER-EMPLOYEE DATA

2.1 Introduction

Part-time employment represents a non-trivial segment of the United States labor market. In June 2015, about 27.6 million individuals in the U.S. were working part-time, accounting for about 18.4% of employment.[1] While the majority of part-time employment is voluntary, about 24.6% is not.[2] Further, since the early 2000s, part-time employment has been on the rise, with almost all of the increase concentrated during the two recessions of the 2000s (Figure 2.1).

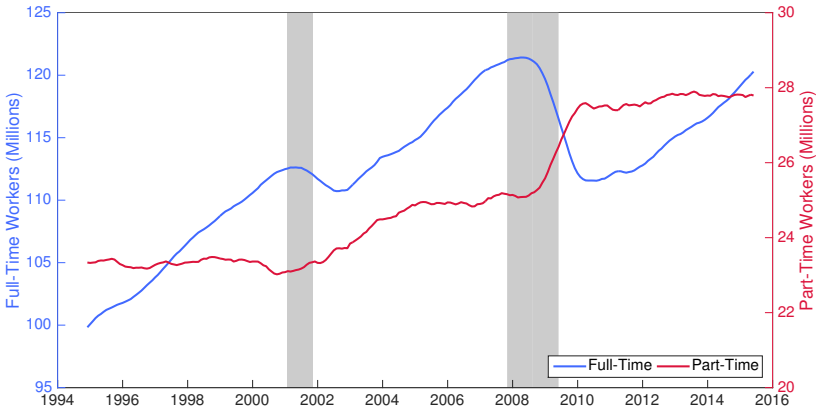


Figure 2.1: Full-Time and Part-Time Workers

Notes: Authors' calculation from the basic monthly files of the Current Population Survey. All time series are MA-smoothed.

[1]This statistics is calculated from the basic monthly files of the Current Population Survey. Part-time employment is defined as workers who work less than 35 hours a week. See Kalleberg (2000) for a summary on the rise in nonstandard employment relationships in the U.S.

[2]There are various reasons why individuals may seek part-time employment. Zabalza et al. (1980) show that part-time work is important in understanding the behavior of workers around retirement. Part-time employment allows for workers to partially withdrawal from the labor market by moving from full-time to part-time. Higgins et al. (2000) analyze how women use part-time employment when balancing work and life, noting that there are both good and bad part-time jobs.

While much of the rise in part-time work after the 2001 recession was in voluntary part-time work, this was not the case after the Great Recession. Figure 2.2 plots the share of part-time employment attributed to involuntary part-time workers. Notice that involuntary part-time work increased drastically during the Great Recession and has remained persistently high. While most of the variation in involuntary part-time work is cyclical, Valletta et al. (2016) find that more persistent features of the labor market, mainly changes in the industry employment shares of aggregate employment and population demographics, are also important in explaining the rise in the incidence of involuntary part-time work.[3] This increase in involuntary part-time work is concerning because of the disadvantages associated with this type of work arrangement for the employees who hold these jobs.[4]

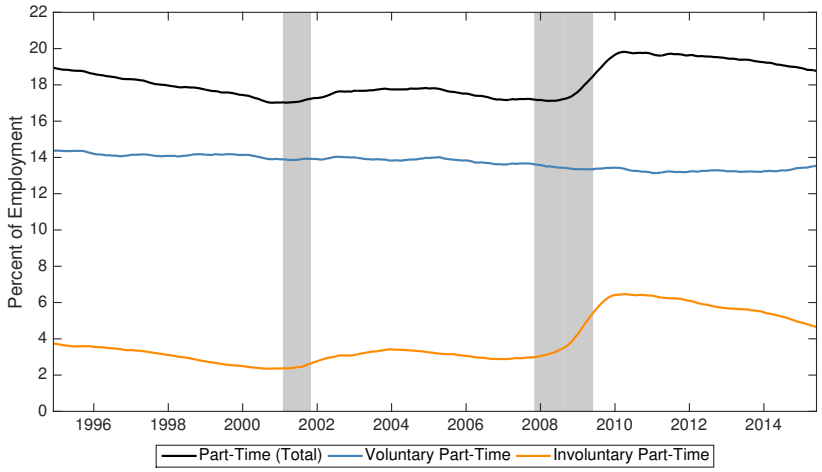


Figure 2.2: Part-Time Employment by Type

Notes: Authors' calculation from the basic monthly files of the Current Population Survey. All time series are MA-smoothed.

[3]See Hijzen and Venn (2011), Valletta and Bengali (2013), Cajner et al. (2014), Canon et al. (2014), Borowczyk-Martins and Lalé (2015), and Valletta and van der List (2015) for further analyses on the recent rise in involuntary part-time employment. There has also been interest on estimating the contribution the Affordable Care Act has on this rise in part-time employment. Even and Macpherson (2015) finds that the impact has been negligible.

[4]For example, Lettau (1997) documents that for the same job, part-time workers are paid a lower wage rate than full-time workers. Further, part-time workers also receive much lower benefits per hour. Aaronson and French (2004) identify a part-time wage effect. They argue that an hours decline causes a wage decline, resulting in a 25% wage penalty for men who cut their work week from 40 to 20 hours. They do not find such an effect for women.

To see how workers transition between full-time work, part-time work, and unemployment, we match the CPS basic monthly files from month-to-month following the algorithm of Shimer (2012). Figure 2.1 plots the transition rates from full-time employment, part-time employment, and unemployment. First, notice that during the Great Recession, both full-time and part-time workers experienced a large increase in the probability of entering unemployment. Further, the Great Recession had a pronounced impact on the movement of full-time workers with a drastic decline in workers staying in full-time employment and an increase in the movement of workers from full-time into part-time employment. Finally, notice that, since the mid-2000s, part-time workers are less likely to transition into full-time employment and more likely to stay in part-time work. Farber (1999a) shows that job losers are significantly more likely than non-job losers to be in both temporary and involuntary part-time jobs. Further, he also finds evidence that temporary and involuntary part-time jobs are part of a transitional process back to regular full-time employment after job loss. Thus, this decline in mobility between full-time and part-time employment may have long-term consequences for workers who lost jobs in this last recession.

While these statistics from the CPS highlight the importance of understanding the growing role of part-time employment in the U.S., it is unclear how firms decide to hire a worker into a full-time position. Do firms have a tendency to increase hours of work from part-time to full-time employment if they need more full-time workers? Or are part-time and full-time positions treated as separate workforces, resulting in firms hiring a new worker to fill a new full-time position? To start such a discussion, we provide the first look at administrative data on hours worked within firms.[5]

[5]These questions are similar to the ones analyzed in studies on the different types of adjustment costs firms face. However, these studies largely rely on firm level data from manufacturers. See, for example, Hamermesh (1989b), Hamermesh and Pfann (1996), Caballero et al. (1997), Hall (2004), Cooper et al. (2004), Cooper et al. (2007), Cooper and Willis (2009), Lee and Mukoyama (2012). As our analysis will show, manufacturing firms are quite different from those in the service industries.

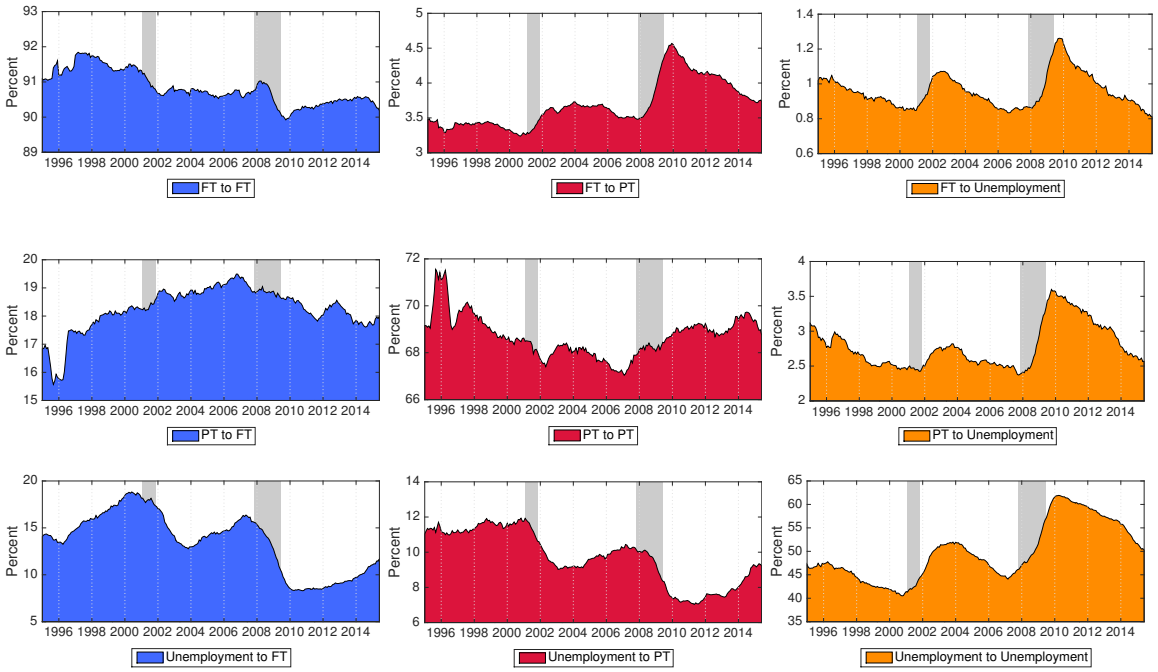


Figure 2.3: Transitions to and from Full-Time and Part-Time Employment

Notes: Authors' calculation from the basic monthly files of the Current Population Survey. Monthly files are matched following the algorithm in Shimer (2012). All time series are MA-smoothed.

First, we document the extent of part-time work across industries. As observed in household surveys, we confirm that part-time work is concentrated in in the relatively low-wage service sectors. Next, we take advantage of the longitudinal nature of our dataset and analyze the prevalence of transitions between part-time and full-time work within the same job.[6] We show that the share of new full-time or part-time jobs that are created due to within job hours changes varies greatly across industries.

The remainder of the chapter is organized as follows. Section 2.2 describes the data and presents some basic descriptive statistics on the dataset. Section 2.3 discusses how the composition of part-time and full-time jobs differs across industries. Section 2.4 defines the transitions of workers into full-time and part-time work used to decompose jobs within a period by the origin employment state of the worker. The section presents results

[6]There is a large body of work that studies worker and job flows. See, for example, Davis and Haltiwanger (1990), Davis and Haltiwanger (1999), Burgess et al. (2000), Davis et al. (2012), Lazear and Spletzer (2012). We add an hours dimensions and analyze how workers flow into part-time and full-time jobs.

on how the relative importance of these different channels on creating new full-time and part-time jobs differs across industries. Finally, section 2.5 concludes.

2.2 Data

To understand how transitions into and out of part-time jobs differs across industries, we use linked employer-employee administrative data from the U.S. Census Bureau. The Longitudinal Employer-Household Dynamics (LEHD) data are a result of the Local Employment Dynamics (LED) partnership between state partners and the U.S. Census Bureau. State partners provide the U.S. Census Bureau with linked employer-employee earnings records from state-administered unemployment insurance programs. At its core, the Employment History File (EHF) is a job-based frame, which consists of an employer identifier, an employee identifier, year, quarter, and gross quarterly earnings for the employee. The data allow us to track jobs, workers, and/or firms over time as well as worker's job-to-job transitions. States also provide characteristics on employers, and the U.S. Census Bureau leverages its own data to create a file of employee characteristics including age, sex, race, ethnicity and education. The LEHD programs uses the microdata to produce public-use data products. Unless otherwise noted, for everything that follows, we use the underlying confidential microdata.[7]

In addition to the core infrastructure files, which make-up the LEHD microdata, four states provide data on gross quarterly hours worked. The state are Washington, Minnesota, Rhode Island, and Oregon. Washington and Minnesota begin reporting hours from the early 1990s, Rhode Island begins reporting quarterly hours to the Census Bureau in 2009Q4, and Oregon begins reporting hours in 2011Q1. Although we have hours data from Washington going as far back as 1990Q1, all results using this hours data must

[7]See Abowd et al. (2009) for the full documentation of the LEHD infrastructure files.

Table 2.1: LEHD States with Employer Reported Hours

State	First YYYY:Q	Last YYYY:Q	Jobs	Jobs (%)	GDP Growth (%)	Δ u rate
			2012:2 QWI Employment		*2009:4 - 2013:4*	
Washington	1990:1	2013:4	2,759,995	2.15	1.79	-3.77
Oregon	2011:1	2013:4	1,592,991	1.24	2.13	-3.60
Minnesota	1998:1	2013:4	2,618,048	2.04	2.13	-3.27
Rhode Island	2009:4	2013:4	436,278	0.34	0.41	-2.17
United States					1.77	-3.0

Notes: Minnesota has a hole in hours reporting from 2002:1 to 2008:1. Data for jobs and share of jobs come the 2016Q1 vintage of the Quarterly Work Force Indicators from the U.S. Census Bureau. GDP growth is from the Bureau of Economic Analysis, Regional Economic Accounts. Changes in unemployment rates are from the Bureau of Labor Statistics Local Area Unemployment Statistics program. Real GDP growth is calculated at an annualized rate.

be from 2009Q4 onwards. This is because, as part of the LED partnership, all results using the LEHD microdata must use at least three states in order to be released. Table 2.1 summarizes the span of hours reporting for each state.

Although a small portion of jobs, the four hours reporting states are broadly representative of the labor market recovery. The fourth and fifth columns of Table 2.1 show beginning-quarter employment for each hours reporting state from the Quarterly Workforce Indicators (QWI) for 2012Q2. Due to the significant seasonality in part-time work, we use the second quarter as the reference quarter to eliminate skewed results due to the holiday season (quarter four), or the summer (quarter three). The hours reporting states unsurprisingly represent a small share of jobs, with Washington being the highest in our sample accounting for 2.15% of national employment. Our sample spans 2009Q4 to 2013Q4, which roughly corresponds to the labor market trough of the Great Recession with the following four years of the recovery. Column 5 of Table 2.1 shows annualized real GDP growth rates for the four states and for the United States. With the exception of Rhode Island, the other three hours reporting states saw GDP growth roughly in line with the U.S. as a whole. Column 6 shows the ensuing decline in unemployment rates in the hours reporting states. Again, Rhode Island is the outlier with the other hours reporting

states seeing slightly larger declines in their unemployment rates compared to the U.S. as a whole.

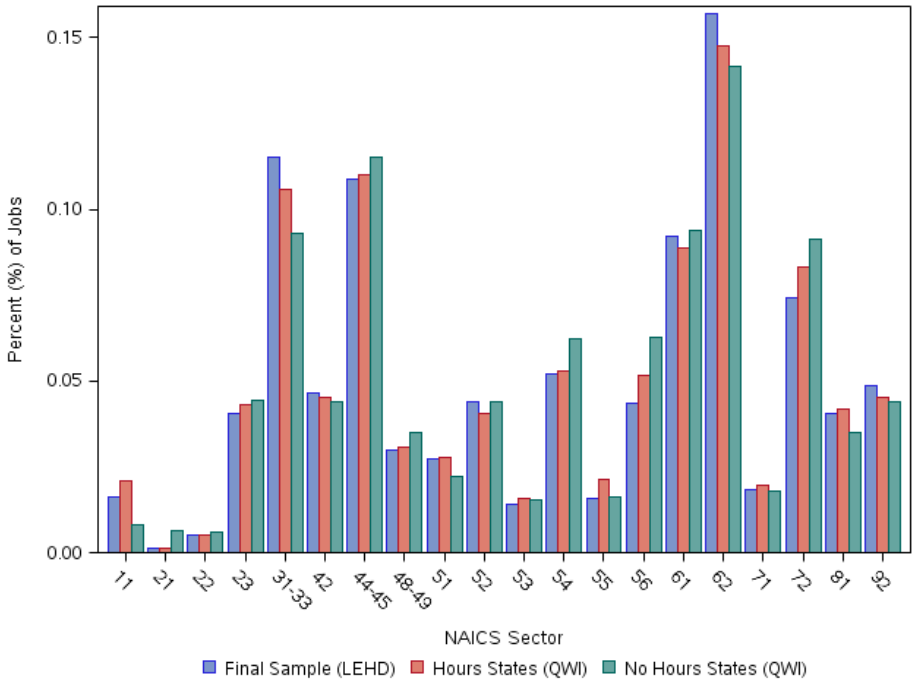


Figure 2.4: Industry Composition of Employment, 2012

Notes: Authors' analysis of Quarterly Work Force Indicators (QWI) from the U.S. Census Bureau. Hours States include Minnesota, Oregon, Rhode Island, and Washington. "No Hours States" include remaining states plus the District of Columbia. Beginning-Quarter Employment is the definition of a job. Final sample is the released LEHD analysis sample with full-quarter employment the definition of a job.

The job composition of the hours reporting states is also broadly reflective of the U.S. job composition. Figure 2.2 shows the share of beginning-quarter jobs from the QWI in each NAICS sector by the pooled hours reporting states and the rest of the United States plus the District of Columbia. Again, we use 2012Q2 as the reference quarter. The distribution of jobs across industries is relatively close to the remaining states. The hours reporting states have a greater percentage of jobs in manufacturing and health care, with the remaining jobs almost uniformly distributed across the remaining NAICS sectors. The result is slightly less jobs in the remaining sectors. Combined with GDP growth and the change in unemployment rates, the composition of jobs indicates that the four hours reporting states are not inconsistent with the labor market composition of the U.S.

in general.

2.2.1 Sample Construction

Unlike the LEHD infrastructure files, which are actively curated and cleaned by dedicated staff, the hours records have not been subjected to the same level of quality assurance. The precise details regarding the cleaning of the hours data prior to constructing our analysis sample are described in an unreleased, confidential Appendix.[8] After this initial data cleaning, we aggregate the hours and earnings information to the PIK-YEAR-QUARTER-SEIN level. This means that a firm in our analysis is defined as an SEIN and a job is a PIK-SEIN pair.[9]

We limit the analysis to stable employees, using the definition of full-quarter employment in Abowd et al. (2009) to select for such jobs. Specifically, let $y_{i,j,t}$ be earnings paid to person i at firm j in period t, where t is a year-quarter. A person i is full quarter employee at a firm j in a period t if he is employed at that same job the period before and the period after:

$$f_{i,j,t} = \begin{cases} 1 & \text{if } y_{i,j,t-1} > 0 \text{ and } y_{i,j,t} > 0 \text{ and } y_{i,j,t+1} > 0 \\ 0 & \text{otherwise} \end{cases}$$

After identifying full-quarter employees at a firm j in period t, we then further restrict the sample to individuals age 18-70 in the period of full-quarter employment.[10] Limiting our analysis to only full-quarter employees is certainly restrictive, especially since part-time jobs may include temporary positions that do not last more than a few months. However, given that we do not know the precise start and end dates of jobs, this restriction means

[8] In general, the quality of the hours data are quite good, and they have an exceptional degree of correspondence to the main infrastructure files.

[9] This only affects records in MN, which are reported at the PIK-YEAR-QUARTER-SEIN-SEINUNIT level. For the other three states, the records are already at the PIK-YEAR-QUARTER-SEIN level.

[10] Note that applying the age restriction after the full-quarter employment restriction means that an individual can be a full-quarter employee at age 18 and at age 70.

we are less likely to mistakenly classify an employee as being part-time due to being only partially employed in a quarter than if we were to include all workers with positive earnings in our analysis.

The cut-off for full-time versus part-time work is important for all results to follow. We denote a job with less than 400 hours in a quarter as a part-time job, and any job greater than or equal to 400 hours a full-time job. We selected the cut-off to minimize the number of full-time jobs wrongly classified as part-time, with the trade-off that some part-time jobs will in some quarters be denoted as full-time. Misclassification arises due to fluctuations in the number of weeks per quarter. This affects workers who are paid either on a weekly or a bi-weekly basis—a large share of employees in the U.S. In March 2013, Burgess (2014) finds that 36.5% of U.S. private business pay their employees bi-weekly and 32.4% pay their employees weekly. For workers paid on a weekly basis, the number of pay periods per quarter fluctuations between 12, 13, and 14. For workers paid on a bi-weekly basis, the number of pay periods per quarter fluctuates between 6 and 7, or between 12 and 14 weeks of paid work. The Current Population Survey denotes anyone who works less than 35 hours per week as part-time. An individual employed at 34 hours a week in a 12 week quarter works 408 hours. Therefore, we set our cut-off for part-time work at 400 hours per quarter. Using 400 hours as a cut-off, we preclude any full-time workers from misclassification, but part-time workers who work the full quarter and with usual weekly hours between 29-33 will fluctuate between full- and part-time depending on the weeks paid in the quarter. In general, there is no "right" cut-off at quarterly frequencies. We find using 400 hours results in a reasonable split between full- and part-time jobs, and we include sensitivity checks for our results.

2.2.2 Descriptive Statistics

Table 2.2 presents aggregate statistics on hours worked for the jobs in our sample from 2011Q1 to 2013Q4. Specially, we report the quarterly counts of firms, workers, and jobs; the total earnings paid and total hours worked across all jobs; and hours worked per job and hours worked per worker. Notice that the worker and job counts do not match and hours worked per worker is always higher than hours worked per job. This is because individuals in our dataset can and do hold multiple jobs in a quarter. While there is a lot of seasonal variation in employment, jobs, and hours that we make no attempt to correct for here, the year-over-year changes in quarter 2 highlight the labor market recovery with employment, jobs, and total hours worked all increasing. Notice however that hours per job and hours per worker have remained fairly constant.

Table 2.2: Descriptive Statistics on Analysis Sample

YYYY:Q	Firms	Workers	Jobs	Total Earnings	Total Hours	Hours per Job	Hours per Worker
2011:Q1	261,000	4,440,000	4,700,000	$60,000	1,920	408	432
2011:Q2	349,000	5,710,000	6,010,000	$74,000	2,610	434	456
2011:Q3	356,000	5,750,000	6,050,000	$77,700	2,660	439	462
2011:Q4	348,000	5,710,000	6,010,000	$77,300	2,610	434	457
2012:Q1	348,000	5,790,000	6,110,000	$78,800	2,580	422	446
2012:Q2	352,000	5,790,000	6,100,000	$76,000	2,650	435	458
2012:Q3	358,000	5,850,000	6,160,000	$77,700	2,630	426	449
2012:Q4	347,000	5,800,000	6,100,000	$80,400	2,660	436	459
2013:Q1	348,000	5,860,000	6,180,000	$79,900	2,600	421	444
2013:Q2	352,000	5,840,000	6,150,000	$77,000	2,670	434	457
2013:Q3	358,000	5,890,000	6,210,000	$79,600	2,660	428	452
2013:Q4	349,000	5,710,000	6,010,000	$79,700	2,630	438	461

Notes: Columns 1, 2, and 3 report the counts of firms (SEIN), workers (PIK), and jobs (PIK-SEIN) in our sample across the four states that report hours. Columns 4 and 5 report the sum of earnings paid ($1,000,000 dollars) and hours worked (1,000,000 hours) across all jobs. Note that earnings are reported in real 2013Q1 dollars, deflated using the CPI. Column 6 reports hours worked per job, which is computed as the ratio of total hours worked and total number of jobs. Column 7 reports hours worked per worker, which is computed as the ratio of total hours worked and total number of workers. All numbers are rounded to three significant digits.

Table 2.3 presents the same statistics as in the aggregate table but by NAICS sector for

2012, averaged over the four quarters.[11] Figure 2.2 compares the industry composition of the jobs in our analysis sample to those reported in QWI for the hours reporting states. In terms of the distribution of jobs across industries, selecting for full-quarter employment does not drastically change the industry composition of our jobs. Notice that hours per job varies a lot across industries from 335 hours per job (Arts, Entertainment, and Recreation) to 556 hours per job (Mining). The five NAICS sectors with the highest and lowest hours per job are:

Sectors with highest hours per job		
556 hrs	21	Mining
516 hrs	31-33	Manufacturing
512 hrs	22	Utilities
493 hrs	42	Wholesale Trade
489 hrs	52	Finance and Insurance

Sectors with lowest hours per job		
335 hrs	71	Arts, Entertainment, and Recreation
340 hrs	82	Other Services
344 hrs	72	Accommodation and Food Services
358 hrs	61	Educational Services
410 hrs	62	Health Care and Social Assistance

Recall that our cutoff for part-time is set at 400 hours per quarter. Notice that almost all of NAICS sectors with the lowest hours per job are below this cutoff, while all the NAICS sectors with the highest hours per job are well above this cutoff. To better understand the role part-time jobs play in each of these industries, we classify our jobs into part-time and full-time jobs.

[11] See Table B.2 for some basic industry statistics from the 2012 Economic Census for reference.

Table 2.3: Descriptive Statistics on Analysis Sample by NAICS Sector (2012)

		Firms		Workers		Jobs		Earnings		Hours		Hrs per	Hrs per
Code	**Industry Title**	*Total*	*Pct*	*Total*	*Pct*	*Total*	*Pct*	*Total*	*Pct*	*Total*	*Pct*	**Job**	**Worker**
11	Agriculture	9,080	2.58	94,600	1.58	99,100	1.62	$765,000	0.98	46,200	1.76	466	488
21	Mining	365	0.10	8,230	0.14	8,260	0.14	$152,000	0.20	4,590	0.17	556	558
22	Utilities	834	0.24	30,400	0.51	30,500	0.50	$652,000	0.83	15,600	0.59	512	513
23	Construction	31,900	9.09	246,000	4.09	248,000	4.05	$3,380,000	4.32	109,000	4.14	442	445
31-33	Manufacturing	17,800	5.08	703,000	11.70	705,000	11.50	$11,400,000	14.60	364,000	13.84	516	517
42	Wholesale Trade	27,700	7.89	285,000	4.75	286,000	4.67	$4,940,000	6.31	141,000	5.36	493	494
44-45	Retail Trade	31,400	8.93	662,000	11.00	669,000	10.90	$5,550,000	7.10	280,000	10.65	418	422
48-49	Transportation and Warehousing	8,490	2.42	182,000	3.02	183,000	3.00	$2,260,000	2.90	82,600	3.14	451	455
51	Information	5,640	1.60	168,000	2.80	168,000	2.75	$4,380,000	5.61	80,000	3.04	475	476
52	Finance and Insurance	13,200	3.76	268,000	4.46	268,000	4.38	$5,450,000	6.96	131,000	4.98	489	489
53	Real Estate Rental and Leasing	12,000	3.40	85,600	1.43	86,900	1.42	$930,000	1.19	36,600	1.39	421	428
54	Professional Services	38,100	10.90	319,000	5.31	320,000	5.23	$6,220,000	7.95	148,000	5.63	464	465
55	Management of Companies	1,160	0.33	98,100	1.63	98,100	1.60	$2,420,000	3.09	47,700	1.81	486	486
56	Administrative Support	17,300	4.94	263,000	4.39	266,000	4.35	$2,530,000	3.24	112,000	4.26	422	427
61	Educational Services	6,140	1.75	554,000	9.23	564,000	9.23	$6,240,000	7.98	202,000	7.68	358	365
62	Health Care and Social Assistance	31,400	8.94	922,000	15.40	959,000	15.70	$11,400,000	14.60	393,000	14.94	410	426
71	Arts, Entertainment, and Recreation	5,370	1.53	111,000	1.84	113,000	1.85	$817,000	1.05	37,900	1.44	335	343
72	Accommodation and Food Services	26,700	7.59	436,000	7.26	454,000	7.42	$2,300,000	2.95	156,000	5.93	344	358
81	Other Services	63,300	18.00	225,000	3.75	247,000	4.04	$1,720,000	2.20	84,100	3.20	340	373
92	Public Administration	2,980	0.85	295,000	4.92	298,000	4.86	$4,070,000	5.21	137,000	5.21	462	465

Notes: Columns 1, 2, and 3 report the counts and the share of firms (SEIN), workers (PIK), and jobs (PIK-SEIN) in our sample by NAICS sectors. Columns 4 and 5 report the total and share of earnings paid ($1,000) and hours worked (1,000) by NAICS sector. Note that earnings are reported in real 2013Q1 dollars, deflated using the CPI. Column 6 reports hours worked per job, which is computed as the ratio of total hours worked and total number of jobs. Column 7 reports hours worked per worker, which is computed as the ratio of total hours worked and total number of workers.

2.3 Part-Time and Full-Time Jobs

As discussed in Section 2.2, we define part-time jobs as those where the employee worked less than 400 hours in a quarter. Specifically, let $h_{i,j,t}$ be hours worked by individual i at firm j in period t. This individual is a part-time employee in period t if he is working less than 400 hours during his period of full-quarter employment. Otherwise, he is a full-time employee.

$$\text{PT}_{i,j,t} = \begin{cases} 1 & \text{if } y_{i,j,t-1} > 0 \text{ and } y_{i,j,t} > 0 \text{ and } y_{i,j,t+1} > 0 \text{ and } h_{i,j,t} < 400 \\ 0 & \text{otherwise} \end{cases}$$

$$\text{FT}_{i,j,t} = \begin{cases} 1 & \text{if } y_{i,j,t-1} > 0 \text{ and } y_{i,j,t} > 0 \text{ and } y_{i,j,t+1} > 0 \text{ and } h_{i,j,t} \geq 400 \\ 0 & \text{otherwise} \end{cases}$$

Table 2.4 presents statistics on part-time and full-time jobs in our final sample of jobs. About 30.55% of jobs each quarter are part-time, which account for about 15.65% of hours worked. Since we restrict our analysis to full-quarter employees, we are going to miss short-term jobs that are likely to be part-time. Therefore, the share of jobs that are part-time reported in Table 2.4 is likely to be a lower bound. Looking at year-over-year changes we see that full-time work, both in terms of the number of jobs and the number of hours, has been increasing. On the other hand, part-time work seems to have plateaued. There was growth in terms of both jobs and hours from 2011Q2 to 2012Q2, but none from 2012Q2 to 2013Q2. Once again, hours per job for both full-time and part-time jobs remains fairly constant.

The above statistics were also computed by NAICS sector in 2012, averaged over the four quarters (Table 2.5). Across industries, there is a large amount of variation in the fraction of jobs that are part-time, from as low as 8.29% (Utilities) to 58.7% (Accommodation and Food Services). Figure 2.5 plots the share of jobs that are full-time and part-time by industry. For all the industry level plots in this paper, the NAICS sectors are organized

Table 2.4: Part-Time vs. Full-Time Jobs in Analysis Sample

YYYY:Q	FT Jobs	PT Jobs	Share PT Jobs	Total FT Hours	Total PT Hours	Share PT Hours	Hours per FT Job	Hours per PT Job
2011:Q1	3,130	1,580	33.5%	1,580,000	342,000	17.8%	505	217
2011:Q2	4,210	1,800	29.9%	2,200,000	405,000	15.6%	523	225
2011:Q3	4,230	1,830	30.2%	2,280,000	379,000	14.3%	539	207
2011:Q4	4,220	1,790	29.8%	2,210,000	403,000	15.4%	523	226
2012:Q1	4,150	1,960	32.1%	2,150,000	427,000	16.6%	518	218
2012:Q2	4,280	1,820	29.8%	2,240,000	412,000	15.5%	523	227
2012:Q3	4,230	1,930	31.4%	2,220,000	411,000	15.6%	524	213
2012:Q4	4,330	1,760	28.9%	2,260,000	398,000	15.0%	522	226
2013:Q1	4,200	1,980	32.1%	2,160,000	439,000	16.9%	515	222
2013:Q2	4,330	1,820	29.6%	2,260,000	411,000	15.4%	522	226
2013:Q3	4,300	1,910	30.8%	2,250,000	405,000	15.2%	524	212
2013:Q4	4,300	1,710	28.5%	2,250,000	382,000	14.5%	523	223

Notes: Columns 1 and 2 report the number of full-time and part-time jobs (1,000 jobs) in our sample. Column 3 reports the share of jobs that are part-time. Columns 4 and 5 report the total hours worked (1,000 hours) in full-time and part-time jobs. Column 6 reports the share of hours worked at part-time jobs. Columns 7 and 8 report totals hours worked per full-time and per part-time job.

and color-coded following the Bureau of Labor Statistics' further aggregation of NAICS sectors into groupings called "Supersectors."[12] The five NAICS sectors with the highest and lowest shares of part-time jobs are listed below.

Sectors with highest share of part-time jobs

58.7%	72	Accommodation and Food Services
51.4%	81	Other Services
51.2%	71	Arts, Entertainment, and Recreation
49.1%	61	Educational Services
36.6%	62	Health Care and Social Assistance

Sectors with lowest share of part-time jobs

8.29%	22	Utilities
8.97%	31-33	Manufacturing
10.0%	21	Mining
10.4%	55	Management of Companies and Enterprises
11.8%	52	Finance and Insurance

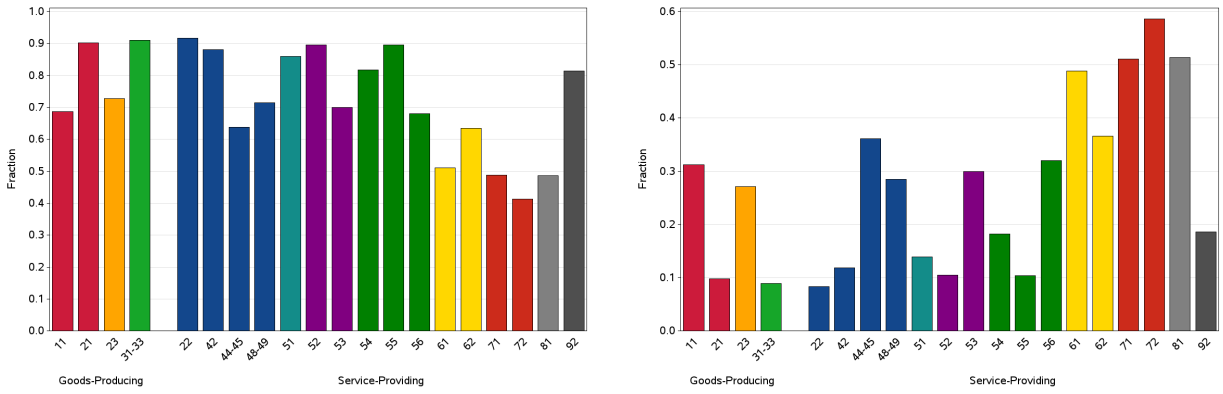
The above results confirm what has already been noted in household surveys, that part-time employment tends to be concentrated in the service sectors. In the next section, we take advantage of the longitudinal structure of LEHD and analyze the work histories of the part-time and full-time employees to shed insight on if and how transitions into part-time and full-time jobs are similar or different.

[12]See Table B.1 for the aggregation. See also `http://www.bls.gov/sae/saesuper.htm`.

Table 2.5: Part-Time vs. Full-Time Jobs in Hours Reporting States by NAICS Sector (2012)

Code	Industry Title	FT Jobs	PT Jobs	Share PT Jobs	FT Hours	PT Hours	Share PT Hours	Hours per FT Job	Hours per PT Job
11	Agriculture	68,200	30,900	31.40%	39,800	6,370	14.1%	581	206
21	Mining	7,450	808	10.00%	4,420	171	3.87%	593	220
22	Utilities	28,000	2,530	8.29%	15,000	596	3.83%	537	235
23	Construction	180,000	67,200	27.30%	94,000	15,300	14.2%	521	229
31-33	Manufacturing	641,000	63,200	8.97%	348,000	15,700	4.31%	542	248
42	Wholesale Trade	252,000	33,800	11.80%	133,000	7,310	5.2%	530	216
44-45	Retail Trade	427,000	241,000	36.10%	223,000	56,800	20.3%	522	235
48-49	Transportation and Warehousing	131,000	52,300	28.60%	70,900	11,700	14.2%	541	223
51	Information	145,000	23,500	14.00%	74,800	5,180	6.48%	516	220
52	Finance and Insurance	240,000	28,000	10.40%	124,000	7,250	5.53%	515	259
53	Real Estate Rental and Leasing	60,800	26,100	30.00%	31,500	5,100	13.9%	518	196
54	Professional Services	262,000	58,400	18.20%	136,000	12,600	8.52%	519	216
55	Management of Companies	87,900	10,200	10.40%	44,900	2,740	5.74%	511	267
56	Administrative Support	181,000	85,100	32.00%	94,800	17,500	15.6%	524	206
61	Educational Services	289,000	276,000	49.10%	145,000	57,400	29.2%	503	214
62	Health Care and Social Assistance	608,000	351,000	36.60%	310,000	82,800	21.1%	510	236
71	Arts, Entertainment, and Recreation	55,200	57,800	51.20%	28,100	9,740	25.8%	510	169
72	Accommodation and Food Services	188,000	266,000	58.70%	94,100	62,000	39.8%	501	233
81	Other Services	120,000	127,000	51.40%	62,500	21,700	25.8%	520	170
92	Public Administration	242,000	55,400	18.60%	126,000	11,300	8.23%	521	204

Notes: Columns 1 and 2 report the number of full-time and part-time jobs by industry. Column 3 reports the share of jobs that are part-time. Columns 4 and 5 report the total hours worked (1,000 hours) in full-time and part-time jobs. Column 6 reports the share of hours worked at part-time jobs. Columns 7 and 8 report totals hours worked per full-time and per part-time job.



(a) Full-Time (b) Part-Time

Figure 2.5: Fraction of Jobs Full-Time and Part-Time by Industry

Notes: The figures plot the fraction of jobs that are full-time and the fraction of jobs that are part-time by industry in 2012. The counts of full-time and part-time jobs are first averaged over the quarters, then shares were computed.

2.4 Transition Dynamics of Part-Time and Full-Time Jobs

To start, we first present the precise definitions of the transitions used to decompose the employees in a period t by how they came to be employed in their part-time or full-time position. Recall that an person i is a stable employee at a firm j in a period t if he is employed for the full quarter:

$$f_{i,j,t} = \begin{cases} 1 & \text{if } y_{i,j,t-1} > 0 \text{ and } y_{i,j,t} > 0 \text{ and } y_{i,j,t+1} > 0 \\ 0 & \text{otherwise} \end{cases}$$

In a period t, stable employees at a firm j can be decomposed by either tracking where they came from or where they are moving to. In particular, we can classify workers by the transitions they made from $t-2$ to $t-1$, how they flowed into stable employment at a firm, or by their transitions from $t+1$ to $t+2$, how they flow out of stable employment at a firm.

Starting with the transitions from $t-2$ to $t-1$, a worker who is stable employed at

a firm j in period t can either be a new hire or a stayer. Specifically, an individual i is a stable new hire at a firm j in period t if he is full-quarter employed in period t, but was not working at firm j in period $t-2$:

$$\text{hire}^{s,\text{in}}_{i,j,t}(t-1|t-2) = \begin{cases} 1 & \text{if } y_{i,j,t-2} = 0 \text{ and } y_{i,j,t-1} > 0 \text{ and } y_{i,j,t} > 0 \text{ and } y_{i,j,t+1} > 0 \\ 0 & \text{otherwise} \end{cases}$$

And an individual i is a stable stayer at a firm j in period t if he is full-quarter employed in period t and was working in period $t-2$:

$$\text{stayer}^{s,\text{in}}_{i,j,t}(t-1|t-2) = \begin{cases} 1 & \text{if } y_{i,j,t-2} > 0 \text{ and } y_{i,j,t-1} > 0 \text{ and } y_{i,j,t} > 0 \text{ and } y_{i,j,t+1} > 0 \\ 0 & \text{otherwise} \end{cases}$$

Similarly, tracking the transitions of workers from $t+1$ to $t+2$, stable workers can either be leavers in period t or stayers. Specifically, an individual i is a stable leaver at a firm j in period t if he is full-quarter employed in period t, but is no longer working for firm j in period $t+2$:

$$\text{leaver}^{s,\text{out}}_{i,j,t}(t+2|t+1) = \begin{cases} 1 & \text{if } y_{i,j,t-1} > 0 \text{ and } y_{i,j,t} > 0 \text{ and } y_{i,j,t+1} > 0 \text{ and } y_{i,j,t+2} = 0 \\ 0 & \text{otherwise} \end{cases}$$

Finally, an individual i is a stable stayer at a firm j in period t if he is full-quarter employed in period t and is also working in period $t+2$:

$$\text{stayer}^{s,\text{out}}_{i,j,t}(t+2|t+1) = \begin{cases} 1 & \text{if } y_{i,j,t-2} > 0 \text{ and } y_{i,j,t} > 0 \text{ and } y_{i,j,t+1} > 0 \text{ and } y_{i,j,t+2} > 0 \\ 0 & \text{otherwise} \end{cases}$$

These stable transitions are summarized in Table 2.6 below.

Define $N_{j,t}$ as employment at firm j in period t, and $\mathcal{N}_{j,t}$ as the set of full-quarter workers i employed at firm j in period t. The above transitions allow us to decompose the stable workforce at a firm j at time t in two ways:

$$\begin{aligned} N_{j,t} &= \sum_{i \in \mathcal{N}_{j,t}} \mathbb{1}\{\text{hire}^{s,\text{in}}_{i,j,t}(t-1|t-2) = 1\} + \sum_{i \in \mathcal{N}_{j,t}} \mathbb{1}\{\text{stayer}^{s,\text{in}}_{i,j,t}(t-1|t-2) = 1\} \\ &= \sum_{i \in \mathcal{N}_{j,t}} \mathbb{1}\{\text{leaver}^{s,\text{out}}_{i,j,t}(t+2|t+1) = 1\} + \sum_{i \in \mathcal{N}_{j,t}} \mathbb{1}\{\text{stayer}^{s,\text{out}}_{i,j,t}(t+2|t+1) = 1\} \end{aligned}$$

Table 2.6: Stable Worker Transitions

	$t-2$	$t-1$	t	$t+1$	$t+2$
$f_{i,j,t}$	——	$y_{i,j,t-1} > 0$	$y_{i,j,t} > 0$	$y_{i,j,t+1} > 0$	——
$\text{hire}^{s,\text{in}}_{i,j,t}(t-1\|t-2)$	$y_{i,j,t-2} = 0$	$y_{i,j,t-1} > 0$	$y_{i,j,t} > 0$	$y_{i,j,t+1} > 0$	——
$\text{stayer}^{s,\text{in}}_{i,j,t}(t-1\|t-2)$	$y_{i,j,t-2} > 0$	$y_{i,j,t-1} > 0$	$y_{i,j,t} > 0$	$y_{i,j,t+1} > 0$	——
$\text{leaver}^{s,\text{out}}_{i,j,t}(t+2\|t+1)$	——	$y_{i,j,t-1} > 0$	$y_{i,j,t} > 0$	$y_{i,j,t+1} > 0$	$y_{i,j,t+2} = 0$
$\text{stayer}^{s,\text{out}}_{i,j,t}(t+2\|t+1)$	——	$y_{i,j,t-1} > 0$	$y_{i,j,t} > 0$	$y_{i,j,t+1} > 0$	$y_{i,j,t+2} > 0$

Notes: Decomposing the stable workers at a firm j in period t based on either inflows from $t-2$ to $t-1$ or outflows from $t+1$ to $t+2$.

where $\mathbb{1}\{\cdot\}$ is the indicator function. Given this decomposition of the workforce, we can rewrite the total labor input at a firm j in period t, $H_{j,t} \equiv \sum_{i \in \mathcal{N}_{j,t}} h_{i,j,t}$, in the following way:

$$\begin{aligned} H_{j,t} &= \sum_{i \in \mathcal{N}_{j,t}} \left[\mathbb{1}\{\text{hire}^{s,\text{in}}_{i,j,t}(t-1|t-2) = 1\} h_{i,j,t} + \mathbb{1}\{\text{stayer}^{s,\text{in}}_{i,j,t}(t-1|t-2) = 1\} h_{i,j,t} \right] \\ &= \sum_{i \in \mathcal{N}_{j,t}} \left[\mathbb{1}\{\text{leaver}^{s,\text{out}}_{i,j,t}(t+2|t+1) = 1\} h_{i,j,t} + \mathbb{1}\{\text{stayer}^{s,\text{out}}_{i,j,t}(t+2|t+1) = 1\} h_{i,j,t} \right] \end{aligned}$$

This decomposition allows us to decompose jobs and hours worked into those attributed to new hires and those attributed to workers who were already employed at the firm.

Using hours and work history information, these transitions can be further decomposed into movements into full-time and part-time employment. Specifically, given the hours worked in time t, new hires can be decomposed into hires into part-time employment and hires into full-time employment. Further, given the longitudinal structure of LEHD, we can also track whether the worker was employed at another firm in the period prior to hiring. If the worker made positive earnings elsewhere in the quarter prior to being hired by firm j, we say the worker was hired from employment. If not, then the worker was hired from nonemployment. Leavers to employment or nonemployment from part-time and full-time jobs are defined similarly. Finally, stayers are workers with two consecutive quarters of full-quarter employment. Given the hours information in both these quarter, we can infer the transition dynamics of these workers between full-

time and part-time employment. The exact definitions of these detailed transitions are in Appendix B.1, however they are also summarized in Table 2.7 below.

Table 2.7: Worker Transitions Supplemented with Hours and Work History Information

	$t-2$	$t-1$	t
$\text{PT}_{i,j,t}$	——	——	$h_{i,j,t} < 400$
$\text{FT}_{i,j,t}$	——	——	$h_{i,j,t} \geq 400$
	Full-Time Employment: $\text{FT}_{i,j,t} = 1$		
$\text{hire}^{s,\text{in}}_{i,j,t}(\text{FT}\vert\text{E})$	$y_{i,j',t-2} > 0, j' \neq j$	——	$\text{hire}^{s,\text{in}}_{i,j,t} = 1$
$\text{hire}^{s,\text{in},u}_{i,j,t}(\text{FT}\vert\text{NE})$	$y_{i,j',t-2} = 0, \forall j'$	——	$\text{hire}^{s,\text{in}}_{i,j,t} = 1$
$\text{stayer}^{s,\text{in}}_{i,j,t}(\text{FT}\vert\text{PT})$	——	$h_{i,j,t-1} < 400$	$\text{stayer}^{s,\text{in}}_{i,j,t} = 1$
$\text{stayer}^{s,\text{in}}_{i,j,t}(\text{FT}\vert\text{FT})$	——	$h_{i,j,t-1} \geq 400$	$\text{stayer}^{s,\text{in}}_{i,j,t} = 1$
	Part-Time Employment: $\text{PT}_{i,j,t} = 1$		
$\text{hire}^{s,\text{in}}_{i,j,t}(\text{PT}\vert\text{E})$	$y_{i,j',t-2} > 0, j' \neq j$	——	$\text{hire}^{s,\text{in}}_{i,j,t} = 1$
$\text{hire}^{s,\text{in},u}_{i,j,t}(\text{PT}\vert\text{NE})$	$y_{i,j',t-2} = 0, \forall j'$	——	$\text{hire}^{s,\text{in}}_{i,j,t} = 1$
$\text{stayer}^{s,\text{in}}_{i,j,t}(\text{PT}\vert\text{PT})$	——	$h_{i,j,t-1} < 400$	$\text{stayer}^{s,\text{in}}_{i,j,t} = 1$
$\text{stayer}^{s,\text{in}}_{i,j,t}(\text{PT}\vert\text{FT})$	——	$h_{i,j,t-1} \geq 400$	$\text{stayer}^{s,\text{in}}_{i,j,t} = 1$

Notes: Decomposing the stable workers at a firm j in period t based on either inflows from $t-2$ to $t-1$ or outflows from $t+1$ to $t+2$ while incorporating information on both hours worked and the worker's employment history.

Before decomposing employment in each industry by all possible origin states, we start by presenting statistics on new hires. In particular, Table 2.8 reports the fraction of new hires that are to full-time jobs and the complementary fraction of new hires that are to part-time jobs. Pooled, new hires are about evenly split between full-time and part-time jobs: 54.6% of new hires are to full-time jobs and 45.5% are to part-time jobs. However, this mix between full-time and part-time jobs among new hires varies quite a bit by industry (Figure 2.6). The sectors with the highest and lowest share of new hires that are to part-time jobs are:

Sectors with highest share of hires to PT			Sectors with lowest share of hires to PT		
71.0%	72	Accommodation and Food Services	12.4%	55	Management of Companies and Enterprises
69.3%	61	Educational Services	13.5%	21	Mining
68.6%	71	Arts, Entertainment, and Recreation	16.3%	31-33	Manufacturing
62.6%	81	Other Serives	16.8%	52	Finance and Insurance
53.5%	44-45	Retail Trade	19.2%	22	Utilities

Notice that the differences between sectors in the fraction of new hires that are part-time jobs is very similar to the fraction of jobs in an industry that are part-time. In other words, the general shape of Figure 2.6 mimics that of Figure 2.5.

Table 2.8: Fraction of New Hires Full-Time and Part-Time by Industry

Code	Industry Title	Full-Time Jobs	Part-Time Jobs
	Aggregate	54.6%	45.4%
11	Agriculture	55.2%	44.8%
21	Mining	86.5%	13.5%
22	Utilities	80.8%	19.2%
23	Construction	64.4%	35.6%
31-33	Manufacturing	83.7%	16.3%
42	Wholesale Trade	79.4%	20.6%
44-45	Retail Trade	46.5%	53.5%
48-49	Transportation and Warehousing	65.4%	34.6%
51	Information	76.7%	23.3%
52	Finance and Insurance	83.2%	16.8%
53	Real Estate Rental and Leasing	58.7%	41.3%
54	Professional Services	73.7%	26.3%
55	Management of Companies	87.6%	12.4%
56	Administrative Support	59.7%	40.3%
61	Educational Services	30.7%	69.3%
62	Health Care and Social Assistance	51.1%	48.9%
71	Arts, Entertainment, and Recreation	31.4%	68.6%
72	Accommodation and Food Services	29.0%	71.0%
81	Other Services	37.4%	62.6%
92	Public Administration	59.1%	40.9%

Notes: Columns 1 and 2 report the percentage of new hires that are full-time and part-time, respectively. Note that each row sums to 100%.

Table 2.9 provides the aggregate decomposition of full-time and part-time jobs into hires and stayers.[13] The percent of jobs that are new hires is about twice as high for part-time jobs (13.3%) than it is for full-time jobs (7.01%). This is consistent with part-time

[13]The transitions to full-time employment are plotted in Figure B.1. The transitions to part-time employment are plotted in Figure B.2.

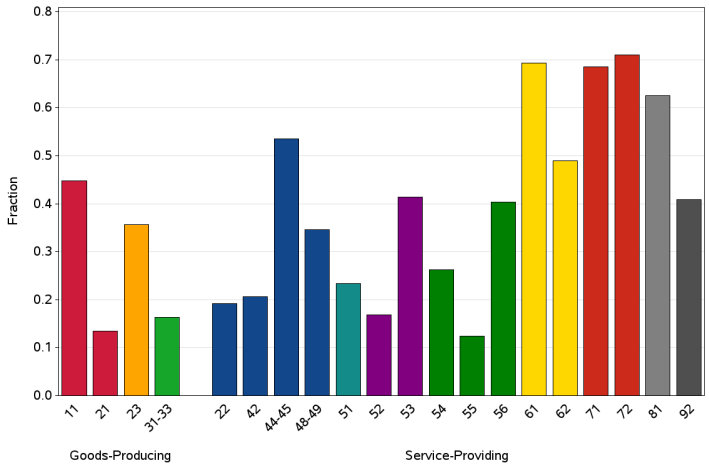


Figure 2.6: Fraction of New Hires that are Part-Time by Industry

Notes: This figure plots the fraction of new hires that are to part-time jobs by NAICS sector.

jobs being more transient and temporary than full-time jobs. This ordering that a higher fraction of part-time jobs tends to be new hires also holds within industries (Table 2.10).

Table 2.9: Transitions of Full-Time and Part-Time Employees

	Full-Time Jobs	**Part-Time Jobs**
Number of jobs	4,250,000 (69.40%)	1,870,000 (30.60%)
% of jobs new hires	7.01%	13.3%
% of jobs stayers	92.99%	86.7%

Notes: Row 1 reports the number of jobs in 2012, averaged over the quarters, that are full-time and part-time. The shares are reported in parentheses. Row 2 reports the percentage of full-time and part-time jobs that are new hires, respectively. Row 3 reports the percentage of full-time and part-time jobs that are stayers. Note that rows 2 and 3 sum to 100%.

Employment can be further decomposed using hours and work history information. See Table 2.10 for the full set of transitions into full-time and part-time work by industry. Pooled, 85.1% of full-time jobs are continuations of an existing full-time job, while 68.6% of part-time jobs are continuations of an existing part-time job. This means that full-time work is a more persistent employment state than part-time work. Once again, this

observation also holds within industry.

Table 2.10: Transitions to Full-Time and Part-Time Employment

				Transitions to FT Jobs						Transitions to PT Jobs			
Code	**Industry Title**	**FT Jobs**		*NE to FT*	*E to FT*	*PT to FT*	*FT to FT*	**PT Jobs**		*NE to PT*	*E to PT*	*PT to PT*	*FT to PT*
	Aggregate	4,250,000	(69.40%)	3.18%	3.83%	7.88%	85.1%	1,870,000	(30.60%)	6.52%	6.74%	68.6%	17.8%
11	Agriculture	68,200	(68.80%)	5.89%	5.82%	10.9%	77.3%	30,900	(31.20%)	9.40%	11.5%	54.4%	24.5%
21	Mining	7,450	(90.20%)	4.90%	3.26%	4.27%	87.6%	808	(9.78%)	7.64%	4.05%	47.8%	40.5%
22	Utilities	28,000	(91.70%)	0.91%	1.45%	3.44%	94.2%	2,530	(8.29%)	3.29%	2.89%	53.9%	39.7%
23	Construction	180,000	(72.90%)	5.45%	5.37%	10.3%	78.8%	67,200	(27.10%)	8.56%	7.51%	55.8%	28.0%
31-33	Manufacturing	641,000	(91.00%)	2.19%	3.17%	3.62%	91.0%	63,200	(8.97%)	5.63%	5.00%	50.8%	38.4%
42	Wholesale Trade	252,000	(88.20%)	2.52%	3.73%	2.99%	90.8%	33,800	(11.80%)	6.29%	5.81%	64.6%	23.1%
44-45	Retail Trade	427,000	(63.90%)	3.41%	3.62%	9.00%	83.9%	241,000	(36.10%)	7.81%	6.47%	68.8%	16.1%
48-49	Transportation and Warehousing	131,000	(71.50%)	3.16%	5.75%	6.68%	84.4%	52,300	(28.50%)	5.22%	6.58%	72.2%	16.0%
51	Information	145,000	(86.10%)	3.16%	3.49%	3.36%	90.0%	23,500	(13.90%)	6.19%	6.31%	65.9%	21.0%
52	Finance and Insurance	240,000	(89.60%)	2.19%	3.59%	3.34%	90.9%	28,000	(10.40%)	4.75%	5.31%	61.0%	28.8%
53	Real Estate Rental and Leasing	60,800	(70.00%)	3.27%	4.93%	4.88%	86.9%	26,100	(30.00%)	6.63%	6.85%	75.1%	11.2%
54	Professional Services	262,000	(81.80%)	3.81%	4.46%	4.29%	87.4%	58,400	(18.20%)	7.20%	6.03%	66.8%	19.8%
55	Management of Companies	87,900	(89.60%)	3.79%	3.51%	4.29%	88.4%	10,200	(10.40%)	4.48%	4.37%	53.0%	38.0%
56	Administrative Support	181,000	(68.00%)	7.36%	8.33%	7.55%	76.7%	85,100	(32.00%)	11.4%	11.1%	61.3%	16.0%
61	Educational Services	289,000	(51.20%)	2.35%	1.74%	21.1%	74.8%	276,000	(48.80%)	5.30%	4.36%	70.0%	20.3%
62	Health Care and Social Assistance	608,000	(63.40%)	2.78%	3.59%	10.5%	83.1%	351,000	(36.60%)	4.65%	5.94%	70.7%	18.6%
71	Arts, Entertainment, and Recreation	55,200	(48.80%)	4.24%	3.82%	9.63%	82.3%	57,800	(51.20%)	7.62%	9.20%	73.4%	9.19%
72	Accommodation and Food Services	188,000	(41.30%)	4.69%	5.45%	15.8%	73.9%	266,000	(58.70%)	8.17%	9.29%	71.0%	10.6%
81	Other Services	120,000	(48.60%)	3.60%	4.14%	7.97%	84.3%	127,000	(51.40%)	5.24%	6.97%	80.2%	7.30%
92	Public Administration	242,000	(81.40%)	1.31%	1.58%	5.32%	91.8%	55,400	(18.60%)	4.42%	4.33%	67.1%	23.9%

Notes: Column 1 presents the counts and share of jobs (in parentheses) that are full-time. Columns 2 through 5 decompose these full-time jobs by their origin employment state: new hires from nonemployment, new hires from employment, transitions from part-time employment, or transitions from full-time employment. Column 6 presents the counts and share of jobs (in parentheses) that are part-time. Columns 7 through 10 decompose these part-time jobs by their origin employment state: new hires from nonemployment, new hires from employment, transitions from part-time employment, or transitions from full-time employment.

What is perhaps more interesting is the relative importance of the extensive and intensive margin when it comes to the newly hired full-time and part-time workers. Table 2.11 presents the fraction of new full-time (part-time) jobs that comes from increases (decreases) in hours for an already existing job and the fraction that comes from new hires either from employment or nonemployment.[14] Pooled, the extensive versus intensive margin of creating new full-time or part-time jobs seem equally important: 52.9% of new full-time jobs come from a within job increase in hours, while 57.3% of new part-time jobs come from a within job decrease in hours. This relative importance of the two margins does vary by industry. For full-time jobs, the fraction of new full-time jobs that are within job hours changes varies from 32.4% (Wholesale Trade) to 83.7% (Educational Services). The sectors with the highest and lowest share of new full-time jobs coming from within jobs increases in hours are:

Sectors with highest share of PT to FT		
83.7%	61	Educational Services
64.8%	92	Public Administration
62.2%	62	Health Care and Social Assistance
61.0%	72	Accommodation and Food Services
59.3%	22	Utilities

Sectors with lowest share of PT to FT		
32.4%	42	Wholesale Trade
32.5%	56	Administrative Support, Waste Management, Rem.
33.6%	51	Information
34.2%	54	Professional, Scientific, and Technical Services
34.4%	21	Mining

The sectors with the highest and lowest share of new full-time jobs coming from within jobs increases in hours are:

Sectors with highest share of PT to FT		
83.7%	61	Educational Services
64.8%	92	Public Administration
62.2%	62	Health Care and Social Assistance
61.0%	72	Accommodation and Food Services
59.3%	22	Utilities

Sectors with lowest share of PT to FT		
32.4%	42	Wholesale Trade
32.5%	56	Administrative Support, Waste Management, Rem.
33.6%	51	Information
34.2%	54	Professional, Scientific, and Technical Services
34.4%	21	Mining

[14] See Figure B.3 for the corresponding plots.

When hiring new full-time workers along the extensive margin, few industries display a higher tendency to hire from nonemployment rather than from employment.[15] However, some industries do seem to have a higher tendency to hire from employment rather than from nonemployment. The top five industries that are more likely to hire from employment when using the extensive margin are:

Sectors with largest difference between E to FT and NE to FT

36.9% – 20.3% = 16.6%	48-49	Transportation and Warehousing
39.3% – 24.0% = 15.3%	52	Finance and Insurance
40.4% – 27.3% = 13.1%	42	Wholesale Trade
37.7% – 25.0% = 12.7%	53	Real Estate Rental and Leasing
35.3% – 24.4% = 10.9%	31-33	Manufacturing

For part-time jobs, the fraction of new part-time jobs that are within job hours changes varies from 35.3% (Arts, Entertainment, and Recreation) to 86.5% (Utilities). The sectors with the highest and lowest share of new part-time jobs coming from within jobs decreases in hours are:

Sectors with highest share of FT to PT

86.5%	22	Utilities
81.1%	55	Management of Companies
78.3%	31-33	Manufacturing
77.6%	21	Mining
74.1%	52	Finance and Insurance

Sectors with lowest share of FT to PT

35.3%	71	Arts, Entertainment, and Recreation
37.4%	81	Other Services
37.8%	72	Accommodation and Food Services
41.5%	56	Administrative Support, Waste Management, Rem.
45.4%	53	Real Estate Rental and Leasing

Unlike hiring for new full-time jobs where some industries have a tendency to hire from employment, new hires for part-time jobs are almost just as likely to come from employment as they are to come from nonemployment. Therefore, job-to-job transitions into full-time employment probably have a job ladder motive while those to part-time jobs do not.[16]

Finally, Table 2.12 reports the hours per job associated with each transition to full-time and part-time work. While new hires and individuals who stay in their employment state (FT to FT or PT to PT) work about the same number of hours per job, workers who change

[15] Mining is the only one with 26.2% of new full-time jobs being hires from employment and 39.4% being hires from nonemployment.

[16] See Topel and Ward (1992) and Farber (1999b).

Table 2.11: New Transitions to Full-Time and Part-Time Employment

		Transitions to New FT			Transitions to New PT		
Code	**Industry Title**	*PT to FT*	*E to FT*	*NE to FT*	*FT to PT*	*E to PT*	*NE to PT*
	Aggregate	52.9%	25.7%	21.4%	57.3%	21.7%	21.0%
11	Agriculture	48.2%	25.8%	26.1%	53.9%	25.4%	20.7%
21	Mining	34.4%	26.2%	39.4%	77.6%	7.77%	14.7%
22	Utilities	59.3%	25.0%	15.6%	86.5%	6.29%	7.18%
23	Construction	48.8%	25.4%	25.8%	63.6%	17.0%	19.4%
31-33	Manufacturing	40.3%	35.3%	24.4%	78.3%	10.2%	11.5%
42	Wholesale Trade	32.4%	40.4%	27.3%	65.6%	16.5%	17.9%
44-45	Retail Trade	56.2%	22.6%	21.3%	53.0%	21.3%	25.7%
48-49	Transportation and Warehousing	42.9%	36.9%	20.3%	57.6%	23.7%	18.8%
51	Information	33.6%	34.9%	31.5%	62.7%	18.8%	18.5%
52	Finance and Insurance	36.6%	39.3%	24.0%	74.1%	13.7%	12.2%
53	Real Estate Rental and Leasing	37.3%	37.7%	25.0%	45.4%	27.8%	26.9%
54	Professional Services	34.2%	35.5%	30.3%	59.9%	18.3%	21.8%
55	Management of Companies	37.0%	30.3%	32.7%	81.1%	9.34%	9.56%
56	Administrative Support	32.5%	35.8%	31.7%	41.5%	28.9%	29.6%
61	Educational Services	83.7%	6.91%	9.34%	67.7%	14.6%	17.7%
62	Health Care and Social Assistance	62.2%	21.3%	16.5%	63.7%	20.4%	15.9%
71	Arts, Entertainment, and Recreation	54.4%	21.6%	24.0%	35.3%	35.4%	29.3%
72	Accommodation and Food Services	61.0%	21.0%	18.1%	37.8%	33.1%	29.1%
81	Other Services	50.7%	26.4%	22.9%	37.4%	35.7%	26.9%
92	Public Administration	64.8%	19.3%	16.0%	73.2%	13.3%	13.5%

Notes: The denominator for transitions to new full-time jobs is the sum of new hires (either from employment or nonemployment) and stayers who transition from part-time to full-time employment. The denominator for transitions to new part-time jobs is the sum of new hires (either from employment or nonemployment) and stayers who transition from full-time to part-time employment.

hours within a jobs do not. In particular, workers who transition from part-time to full-time within a job work slightly fewer hours per job than do the rest of full-time workers. And workers who transition from full-time to part-time within a job work slightly many more hours per job than do the rest of part-time workers. Consistent with other studies on hours adjustments, between job hours changes (new hires) are much higher than within job hours changes.[17]

[17]See Altonji and Paxson (1986, 1988).

Table 2.12: Average Hours per Job for Transitions to Full-Time and Part-Time Employment

				Transitions to Full-Time Jobs							Transitions to Part-Time Jobs					
Code	**Industry Title**	**All Jobs**	**FT Jobs**	*Hires*	*NE to FT*	*E to FT*	*Stayers*	*PT to FT*	*FT to FT*	**PT Jobs**	*Hires*	*NE to PT*	*E to PT*	*Stayers*	*PT to PT*	*FT to PT*
	Aggregate	430	522	520	518	522	522	483	526	220	202	198	206	223	203	303
11	Agriculture	466	584	589	584	594	583	545	588	206	190	179	202	210	185	268
21	Mining	556	594	611	604	615	592	560	594	212	204	193	209	213	193	238
22	Utilities	512	537	529	529	529	537	493	539	236	181	181	181	240	170	335
23	Construction	442	521	534	525	542	520	496	523	228	211	208	214	231	207	278
31-33	Manufacturing	516	542	543	540	546	542	514	544	248	225	227	224	250	211	304
42	Wholesale Trade	493	530	528	530	526	530	506	531	216	204	206	202	218	192	292
44-45	Retail Trade	418	522	512	513	511	522	485	526	235	233	231	235	235	220	304
48-49	Transportation and Warehousing	451	541	535	527	551	542	505	545	223	213	212	214	224	209	292
51	Information	475	516	513	516	510	516	494	517	220	199	183	215	223	196	312
52	Finance and Insurance	489	515	515	512	521	515	489	516	259	245	256	233	261	234	319
53	Real Estate Rental and Leasing	421	518	521	520	523	518	482	520	196	187	180	195	197	181	304
54	Professional Services	464	519	518	516	520	519	494	520	217	200	197	203	219	195	300
55	Management of Companies	486	511	522	515	529	511	482	512	268	232	231	233	272	226	336
56	Administrative Support	422	524	516	515	518	526	493	529	205	199	194	203	207	187	284
61	Educational Services	358	501	498	496	500	501	471	510	208	134	130	137	216	202	268
62	Health Care and Social Assistance	410	510	504	503	506	511	470	516	236	214	210	219	239	215	331
71	Arts, Entertainment, and Recreation	335	510	504	502	506	510	473	514	169	167	153	183	169	152	306
72	Accommodation and Food Services	344	502	491	491	491	503	460	512	233	227	226	229	234	221	326
81	Other Services	340	520	521	517	526	520	479	524	170	165	151	184	171	159	310
92	Public Administration	462	521	513	515	511	521	480	523	204	159	146	172	208	168	322

Notes: Hours per job is calculates as total hours worked associated with a particular transition divided by the total number of jobs involved in a particular transition.

Given these differences in hours worked per job for intensive versus extensive margin changes, addition analysis would need to be done to ensure that the intensive margin changes are not due to changes in classification. However, the differences are quite large and hours worked per job for these intensive margin changes are around 100 hours above and below the cutoff. This is a fairly substantial number of hours worked in a quarter, making it unlikely that all of these intensive margin changes are due to misclassification.

2.5 Conclusion

This paper uses administrative data on hours worked from LEHD to study how workers transition into part-time and full-time jobs. We confirm that part-time jobs are particularly concentrated in relatively low-paying service sectors, while full-time jobs are concentrating in high-paying sectors like manufacturing. While part-time jobs face higher turnover than full-time jobs, not all part-time work appears to be temporary hires. In some industries, the most likely channel through which workers enter full-time employment is through part-time work. While in others, workers are more likely to be new hires. Thus, this paper highlights how worker transitions into full-time and part-time work differs greatly across industries.

CHAPTER 3

TOTAL ERROR AND VARIABILITY MEASURES WITH INTEGRATED DISCLOSURE LIMITATION FOR QUARTERLY WORKFORCE INDICATORS AND LEHD ORIGIN DESTINATION EMPLOYMENT STATISTICS IN ONTHEMAP

3.1 Introduction and Summary

We compute the first comprehensive estimates of total variation for two Longitudinal Employer-Household Dynamics (LEHD) products from the U.S. Census Bureau: the Quarterly Workforce Indicators (QWI), which are public-use tables displayed in QWI Explorer, and the workplace-based LEHD Origin-Destination Employment Statistics (LODES), which are the public-use tables displayed in OnTheMap (OTM) when a workplace report is requested. These labor market indicators are produced from a comprehensive integrated administrative record system known as the LEHD Infrastructure File System, which is based primarily on the linkage between employers and employees provided by state-regulated unemployment insurance (UI) wage records. The theoretical universe to which these records correspond is all statutory jobs in the economy – private and public (excluding federal employees).[1] There is also a benchmark census of all such jobs in the universe: the Quarterly Census of Employment and Wages (QCEW) from the Bureau of Labor Statistic (BLS). We use this census, which is also integrated into the LEHD Infrastructure File System as the finite population that the QWI and LODES tabulations estimate. In principle, the published indicators are subject to errors from coverage, sampling, edit, and imputation. By addressing all of these sources of error in our assessment of total variability, we have created the first comprehensive total quality measures for these data.

[1] Although federal employees are now covered in both QWI and LODES, they are excluded from this evaluation.

Coverage errors are addressed in two ways. First, each wage record is linked to the associated employer record from the putative universe of employers (QCEW). When there is a link, estimated employment from the two sources is compared. A tentative weight is constructed to adjust the LEHD Infrastructure File System estimate of employment. When there is not a link, an entity is added to the LEHD infrastructure version of the QCEW, called the Employer Characteristics File (ECF), to account for this absence. At the end of the processing, a final weight is computed that benchmarks all employment to the BLS published state-level employment totals for the same universe. The effect of this procedure is to transmit the coverage errors into the edit and imputation procedures used to complete the firm level tabulation variables when there is a linkage failure in the data integration. Details of these record-linkage procedures are discussed in Abowd and Vilhuber (2005), Benedetto et al. (2007), and Abowd et al. (2009).

Every job in the universe must have completed data for all the publication variables. The LEHD Infrastructure File System has a fully-integrated collection of probability models that generate multiply-imputed values for all missing data items in the system. Most details are supplied in Abowd et al. (2009) – in particular, the models for imputing missing demographic and workplace characteristics.[2] The system uses the methods first proposed by Rubin (1987) and expanded in Little and Rubin (2002) for analyses using multiply-imputed missing data. The total variance statistics described in this paper are based on specially adapted versions of the Rubin measures generated using the approved QWI disclosure avoidance method: input noise-infusion as described in Abowd et al. (2009) and Abowd et al. (2012).

Users of these total variability measures have several options. The measures are intended to provide the information needed to construct approximate confidence intervals at all levels of stratification for five key publication statistics: total employment,

[2] Abowd et al. (2009) does not document the replacement to the demographic variable imputation methods that were incorporated in 2010. Those methods are documented in Appendix C.1.

beginning-quarter employment, full-quarter employment, total payroll, and average monthly earnings of full-quarter employees. We give detailed guidance on how to use our results to calculate confidence intervals for arbitrary published employment totals and earnings.

The Rubin measures are also designed to summarize the extent to which the variability due to the edit and imputation procedures, as distinct from the variability due to sampling in the underlying data, contributes to total variation. Total variability consists of both between variance generated primarily by edit and imputation, and within variance, which consists to some extent of variability due to sampling. However, the sampling variance is small since in principle we should have the population of firms and jobs.

We also distinguish and account for variability due to sampling and structural zeros. In the language of Bishop et al. (1975), a structural zero occurs whenever there is no reported activity in a cell – that is, no business exists in the cell – and a sampling zero occurs when the cell is at risk to have positive employment (because a business exists) but does not. We treat the probability that a job will be classified in a particular detailed category of the publication tables as potentially random within a fixed population of state jobs. This set of assumptions allows us to model the equivalent of sampling variability as if it were generated by a particular multinomial model.

All five indicators we study are published every quarter in the QWI, stratified by ownership, sub-state geography, detailed industry, worker age, gender, race, ethnicity, and education. The publication tables also cross-classify many of these stratifiers. Beginning-of-quarter employment is the primary tabulation variable in LODES for display in OnTheMap, which is released annually with many of the same stratifiers as in the QWI and sub-state geography down to the block level. Constructing measures of total variability for these indicators is complicated by three related factors. First, the QWI and LODES are produced in separate production streams although they share the core LEHD Infras-

tructure File System and, therefore, are subject to the same sources of variation. Neither production stream saves all the inputs required to calculate total variability. Second, the QWI are revised quarterly, and revised indicators are released for the complete history of the series. Third, the workplace-based statistics produced by the LEHD program for both QWI and LODES/OTM use a confidentiality protection system based on input noise-infusion that constrains the calculation of total variability measures and complicates the release of these measures in a user-friendly format.

Because the QWI production system does not store the implicate threads needed to compute the total variability statistics, the analysis in this paper is based on a re-creation of the production statistics from the research files corresponding to a particular vintage of the QWI. The research code does not exactly match the production code. In particular, there are discrepancies in the counts between the production and research values of the statistics for which we compute total variability measures. Even if the research code did exactly reproduce the publication statistics from one release of QWI, the next quarter's release would not agree exactly for most of the historical data because of the continuous-revision design of QWI. The user must take care when calculating confidence intervals for the published values using the total variability measures tabulated here. There are two available strategies, both of which are discussed in this paper. The user can download a table of total variability measures with the same structure as the tabulations for which confidence intervals are required. In this case, there will be some discordance between the value of the indicator found in the publication tables and the value that was used to calculate the total variability measures. We document when these discrepancies can be important: unsurprisingly, mostly for cells with small tabulation counts. We also provide detailed tables that can be used directly to construct approximate confidence intervals.

Overall, these comprehensive measures of the total quality of QWI and LODES tabulations for five critical variables provide substantial evidence that the system is producing

reliable data. This summary discusses the qualitative results for the main employment indicator used by both QWI and LODES/OTM, beginning-of-quarter employment.

Both QWI and LODES/OTM were designed to allow detailed sub-state geography and industry tabulations. Such a system, of necessity, must be robust to the presence of many cells with very small tabulations and many zeros. We document that the vast majority of zeros result from no reported activity, meaning that the value is exactly zero and is treated as a structural zero. Since QWI and LODES are population tabulations, structural zeros have no variability, which is imposed in our analysis. Some zeros are estimated, and those zeros have total variability. Cells with small published employment totals (for any of the employment measures) do have substantial estimated total variability.

The smallest tabulation values (cells containing counts of one or two) often have 90% confidence intervals of less than plus or minus one, so that they sometimes include zero and three. These values are usually suppressed in QWI but they are released in LODES/OTM. The suppression in QWI is justified because the full hierarchical tabulation is published, reducing the need for custom aggregations; however, QWI users and QWI Explorer, the Census Bureau's own analysis tool for these data, do generate custom tabulations. These custom tabulations must populate the cells with suppressed items using some algorithm. There are no suppressions in LODES/OTM, which completes the data using a synthetic data model based on the posterior predictive distribution of small cell counts (one or two) within a given tract stratified by most of the variables for which LODES tabulations are published. Regardless of the model used, there is still substantial uncertainty in these small tabulations, as our results confirm. Almost all 90% confidence intervals are tighter than the interval zero to five, while the vast majority are less than plus or minus two. Publication of these small tabulations in spite of their substantial relative uncertainty is justified by the flexibility they allow for generating custom tabulation areas, most of which end up with much larger estimated employment totals. These cus-

tom tabulations would be substantially biased by using zero as the estimate when the publication value of a component is suppressed.

For cells where the tabulations are in the range of three to nine, our results indicate that the 90% confidence interval is rarely wider than plus or minus three, and for most tables is less than plus or minus one. For cells where the tabulations are in the range of 10 or more, it makes more sense to summarize the results in terms of percentage variation; i.e., use the coefficient of variation implied by the total variability measure and the estimated count. For tabulations in the range of 10 to 99 jobs, the 90% confidence intervals are rarely larger than plus or minus 25% and are usually in the range of plus or minus 10% to 25%. For tabulations in the range 100 to 999, the widest 90% confidence intervals are plus or minus 20%, and the vast majority of cells in this range have confidence intervals of plus or minus less than 10%. For the largest tabulation areas, 1,000 or more, the widest 90% confidence intervals are approximately plus or minus 5%, and the intervals are usually in the range of plus or minus 1.5%.

The other dimension along which we assess the total variability is the Rubin missingness ratio, which quantifies the proportion of the total variability that arises from the multiple imputation procedures. This is also known as "fraction of missing information" as in Little and Rubin (2002). The complement of the Rubin missingness ratio measures the proportion of total variability that it is due to sampling and other intrinsic sources of randomness in the indicator; that is, the proportion of total variability that would remain if no records required any edits or imputation. As we noted above, the edit and imputation procedures used in QWI and LODES/OTM also induce variability due to sub-state coverage errors.

The Rubin missingness ratio provides a reasonable way to assess the effects of data edits and imputations for both demographic characteristics (age, gender, race, ethnicity, and education) and workplace characteristics (industry and county). When age and gender

are the only two stratifiers used in the publication table, missing data account for about 44% of total variability. When race and ethnicity are the only two stratifiers, missing data account for between 80% and 95% of total variability. When gender and education are the only stratifiers in the publication tables, missing data account for over 95% of total variability. When workplace industry and county are the only stratifiers in the publication tables, missing data account for between 0% and 80% of total variability. It is important to note that even when the Rubin missingness ratio is large, the 90% confidence intervals implied by the total variability measure remain as summarized above. The missingness ratios are a guide to where improvements in the data quality either through the use of measured data from other sources or through better imputation algorithms can reduce total variability the most.

The remainder of this chapter is organized as follows. Section 3.2 provides background on the missing data problem, and the methods for multiply imputing worker and establishment characteristics. Section 3.3 provides formal models for estimating total variability and its associated components in a manner that is fully consistent with the required disclosure avoidance procedures. To the best of our knowledge, these formulas have never been derived or published before. Section 3.4 discusses the detailed results and provides guidance for computing confidence intervals. Section 3.5 concludes.

3.2 Background on QWI, LODES, and the Multiply Imputed Characteristics

The QWI are a public-use data product of the U.S. Census Bureau. Every quarter, local labor market statistics are released by worker demographics, workplace geography, and other employer characteristics. Unlike many labor force statistics derived from surveys of

workers or employers, the QWI are produced from job-based administrative data, where a job is the link of a statutory employee to a statutory employer. This linkage allows the QWI to provide tabulations of labor force statistics by worker and employer characteristics, such as county employment by firm size and gender. In addition, the unique identifiers for the employer and worker allow the QWI to tabulate longitudinal statistics, such as hires, separations, and turnover.

The LODES are similar to the QWI in that they originate from the same job-based frame. However, the LODES data provide geographic detail for both place of work and place of residence, but only release a subset of the labor force statistics in the QWI, and are published annually with statistics derived using the first day of the second quarter of the year (April 1st) as the reference date. The core employment variable, beginning-of-quarter employment, called B below, is used for both the QWI and LODES tabulations.[3]

The QWI and LODES are based on the LEHD Infrastructure File System. The original production version of this system is documented in Abowd et al. (2009). The LEHD infrastructure files are made possible through the Local Employment Dynamics state-federal cooperative agreement where participating states provide the U.S. Census Bureau quarterly extracts of earnings records from their respective UI systems as well as an extract from the QCEW, as specified by a similar federal-state cooperative arrangement between the states and the BLS.

The UI earning records are used to construct a job-based frame for the QWI and LODES. An in-scope job occurs when a worker produces at least one dollar of UI-covered earnings at a non-federal establishment in a given quarter. The LEHD Infrastructure File System then combines this information with additional survey and administrative data to derive individual characteristics such as age, gender, place of birth, race, ethnicity, and

[3]Publication tables for the QWI can be found here: `http://qwiexplorer.ces.census.gov/`. Publication maps for LODES/OTM can be found here: `http://onthemap.ces.census.gov/`.

education, as well as establishment characteristics, such as workplace address and North American Industrial Classification System (NAICS) codes. The LEHD Infrastructure File System was developed using model-based edit and imputation procedures. Every missing data element has been multiply-imputed using an integrated set of models described in Abowd et al. (2009). There are ten implicates for every missing item. Implicates are denoted by $l = 1, \ldots, L$. The missing data models for most of the variables used in this paper, including birth date, gender, race, ethnicity, education, workplace geography, workplace NAICS, firm age, and firm size, have been substantially improved and modified since the 2009 article was written. Because the LEHD Infrastructure File System is rebuilt every quarter from all historical records, the analysis in this paper incorporates all of those model improvements.

The LEHD program receives unemployment insurance records from states without any individual characteristics. The individual characteristics are added to the LEHD data from a variety of Census Bureau surveys and federal administrative data. The five multiply-imputed worker characteristics are birth date, sex, race, ethnicity, and education. The variables are imputed into discrete categories and the imputation process proceeds starting with variable(s) having the least amount of missingness, taking advantage of what is commonly known as a monotone missing data pattern. At each stage, imputations from the earlier rounds are used in the current stages imputation model. First, missing birth date and sex are imputed followed by missing race and ethnicity. Missing education is imputed last. Appendix C.1 contains detailed documentation of the individual characteristics imputation.

In addition to worker characteristics, a separate process imputes the establishment for each job spell in the LEHD data. States send the linked employer-employee data at the employee-firm-state level. In addition, the states send a list of all known establishments owned by the firm within a state. This list includes establishment characteristics such as

industry and geography, as well as the employment counts at each establishment within a quarter. However, with the exception of Minnesota, explicit identifiers linking an employee to an establishment do not exist. In order to produce labor market statistics for detailed industries and geographies, the link allocating a worker to an establishment is multiply imputed.[4]

The QWI and the workplace component of LODES are confidentiality protected using a noise-infusion method applied to the underlying micro-data. Every establishment (identifiers: SEIN, SEINUNIT) in the database has been assigned a unique fuzz factor, δ_j, where j indexes establishments that satisfy the conditions documented in Abowd et al. (2009, 2012). The method for applying this fuzz factor to the publication statistics depends upon whether the publication statistic is based on a magnitude (including counts for an establishment), ratio, or other more complicated statistic. In addition, small magnitude values in the QWI are suppressed with the flag "5: Does not meet Census Bureau publication standards" and significantly distorted publication values are labeled with the flag "9: Significantly distorted." In LODES, values that would be suppressed in QWI are synthesized using an approved probability model that is based on the posterior predictive distribution of the suppressed values conditional on tract-level establishment employment data.

The total variability statistics described in this paper apply to data for all private employers and the current all-employer category in the QWI and LODES data, which excludes federal employees. Statistics that include only federal employees are covered by a different protection procedure. Statistics that aggregate all-employer data (excluding federal employment) with federal employment data must combine the two types of data

[4]The data from Minnesota is used to fit a hierarchical Bayesian model of establishment assignment. The probability of an employee working at a given establishment is estimated in this hierarchical structure with the first part conditioning on the employment size of an establishment, and the second part conditioning on the distance between an employee's residence and an establishment. The model is fit jointly on each of three firm size categories, and the estimated model parameters are used to generate 10 draws of feasible establishments for each job. For further details see Abowd et al. (2009).

from their respective public-use releases.

We extend the QWI noise-infusion methods to cover the protection of the Rubin total variance measure for statistics based upon multiply-imputed missing data. This measure combines the conventional quality measure for published statistics – the design-based sampling variance, corrected for ex post departures from design and finite populations – and a measure that captures the contribution of the model-based missing data imputation procedures: the between-implicate variance of the publication statistic.

3.3 Noise-Infusion Protected Total Variance Measures

This section derives the formulas for noise-infusion protected Rubin total variance measures. To the best of our knowledge, these formulas have never been derived or published before. We restrict our analysis to five core labor force statistics published in the QWI:

- Beginning-of-quarter employment, B, which is equal to the sum of all workers who had positive earnings at an establishment in the current quarter as well as the previous quarter.
- Full-quarter employment, F, which is defined as the sum of all workers who had positive earnings at an establishment in the current quarter in addition to the previous and subsequent quarters.
- Average monthly earnings of full-quarter employees, Z_W3.
- Total flow-employment, M, defined as the sum of all workers who have positive earnings at an establishment at any time in the quarter.
- Total payroll, $W1$, which is the total earnings earned by workers in a quarter.

Beginning-of-quarter employment for quarter 2 (April 1-June 30) is also the primary tabulation variable in LODES/OTM.

The relevant population is a state.[5] At the state level, the QCEW measure of all employment (excluding federal workers) is considered the population. Quarterly weights for the QWI benchmark B to the QCEW month 1 employed population. All statistics defined below are calculated for a given state-year-quarter. Similar to the actual QWI, total variability statistics are produced for the period beginning in 1990, quarter 1 (1990:1). The total variability measures discussed in this paper refer to the QWI release labeled R2014Q4, which covers 1990:1 through 2012:1. All states except Massachusetts, North Carolina, and Colorado are included in the R2012Q4 release. Similar to the actual QWI, total variability statistics are produced for the period beginning in 1990, quarter 1 (1990:1). The total variability measures discussed in this paper refer to the QWI release labeled R2014Q4, which covers 1990:1 through 2012:1.[6]

We adopt, without modification, the noise-infusion methodology described in Abowd et al. (2009) and elaborated in Abowd et al. (2012) to which the reader is referred for more details. The system adds multiplicative noise to tabular output produced from the LEHD Infrastructure File System. The multiplicative noise factors for each establishment are drawn from a two-sided symmetric ramp distribution centered at the value one. The draws from the distribution distort the original input by at least a minimum percentage, and by no more than a maximum percentage. Both of these values are Census confidential. This system is a substantial generalization of the method originally developed by Evans et al. (1998). As applied in the production of the QWI and LODES/OTM, the release statistics are dynamically consistent – the same noise factor is used for an establishment in every quarter of data.

The system also provides protection to employers as well as establishments – all establishments for the same employer within a given state have noise distortion factors on

[5]For simplicity, we include Washington, D.C. when we say "state."

[6]Refer to the table here `http://download.vrdc.cornell.edu/qwipu/starting_dates.html` for the exact start dates for each state. The estimated total variability measures described in this paper can be downloaded here: `http://doi.org/10.3886/E100590V1`.

the same side of unity. The system can provide protection to magnitude measures (the only problem considered by Evans et al. (1998)), ratios, and differences. Employment counts within demographic categories are treated as magnitudes. The protection method for ratios requires that the publication tables include either two magnitudes (e.g., total employment and total payroll) or one magnitude and one ratio (e.g., total employment and average quarterly earnings).[7] We use the ratio form of the QWI noise-distortion protection below.

Multiplicative noise infusion provides confidentiality protection in the following formal sense. The originally reported values of the tabulation variables are never used in the formation of the magnitudes (establishment-level counts) and ratios that are tabulated. Tabulations based upon a small number of establishments (at the limit one) or a small number of employees (at the limit one) contain uncertainty induced by the distribution of the noise factor. This uncertainty limits a users ability to infer attributes to within a range that is confidential. Finally, the physical location of a workplace is not treated as confidential because it is defined as the location where the employee must report for work, and is therefore public.

3.3.1 Population Definitions

To calculate the components of total variance, every quarter we require estimates of the total population, N_{WB}, and the total sample size, N_{UB}. To be consistent with the overall data protection scheme, we must calculate these from the fuzzed data as

$$N_{WB} = \sum_{\forall j} B_j^U w_j \delta_j \equiv \sum_{\forall j} B_j^* \quad \text{and} \tag{3.1}$$

$$N_{UB} = \sum_{\forall j} B_j^U \delta_j \equiv \sum_{\forall j} B_j^{U*} \ , \tag{3.2}$$

[7] We do not use the protection method for differences in this paper.

where B_j^U is the unweighted establishment-level beginning-quarter employment for establishment j, w_j is the QWI establishment weight, δ_j is the unique QWI establishment fuzz factor, B_j^* is the fuzzed-weighted establishment-level count, and B_j^{U*} is the fuzzed-unweighted count establishment-level count. Summing over all firms gives us estimates of N_{WB} and N_{UB} (excluding federal establishments). N_{WF}, N_{UF}, F_j^*, F_j^U, and F_j^{U*} are defined similarly for full-quarter employment, as well as N_{WM}, N_{UM}, M_j^*, M_j^U, and M_j^{U*} for total employment. The population estimate N_{WB} has been benchmarked to the QCEW month-1 employed population via the QWI weights. This procedure is also discussed in Abowd et al. (2009). There is no QCEW population count for full-quarter employment nor total employment. However, N_{WF} and N_{WM} are treated here as the appropriate estimate of the population total for F and M, respectively. Since Z_W3 is computed over the same set of input records as F, its fuzzed-weighted and fuzzed-unweighted population and total sample counts are identical to N_{WF} and N_{UF}. $W1$ is calculated using earnings for all workers, thus, N_{WM} and N_{UM} are the correct population and sample size for this statistic.

In principle, for all the missing data models, there should not be any between-implicate variance in N_{WB}, N_{UB}, N_{WF}, N_{UF}, N_{WM} and N_{UM} because missing records are corrected using the weights and only missing items on actual records are imputed. Therefore, it should not make any difference which implicate is used to compute these population and sample totals. We computed population totals separately for each implicate and attempted to verify the absence of between-implicate variation in the total fuzzed-weighted and fuzzed-unweighted counts. In practice, there is a small amount of between-implicate variance in the population totals – less than 0.04% for B and less than 0.03% for F as measured by the coefficient of variation. This result is tabulated in detail in Appendix Table C.11 for the beginning-of-quarter population and in Appendix Table C.12 for the full-quarter population. The between-variance measures are also computed for each establishment type (private and all, excluding federal). These results are also displayed in Appendix Tables C.11 and C.12. Between-implicate variation in the sub-population totals

is consistent with the benchmarking but is also minimal.

3.3.2 Total Variability Models for B, F, and M

Let B_k be any cross-classification of beginning-of-quarter employment such that $N_{WB} = \sum_{\forall k} B_k$. For each implicate l, the fuzzed-weighted count for category k is computed as

$$B_k^{(l)*} = \sum_{(i,j)\in\{\text{def } k\}} b_{i,j}^{(l)} w_j \delta_j \tag{3.3}$$

where $b_{i,j}^{(l)}$ is the LEHD infrastructure indicator variable that defines person i as a beginning-of-quarter employee of establishment j in the l^{th} implicate (implicitly, for date t), {def k} is the set that defines membership in category k for the pair (i, j), and w_j is the QWI weight for establishment j. $F_k^{(l)*}$ and $M_k^{(l)*}$ are defined comparably using the LEHD infrastructure indicator variables $f_{i,j}^{(l)}$ and $m_{i,j}^{(l)}$, respectively, and the same weight and fuzz factor as in the equation for $B_k^{(l)*}$.

For each implicate, the estimated proportion of N_{WB} represented by $B_k^{(l)*}$ in each cell k is

$$p_k^{(l)*} = \frac{B_k^{(l)*}}{N_{WB}} \ . \tag{3.4}$$

The estimated count in cell k can be rewritten as

$$B_k^{(l)*} \equiv c_k^{(l)*} = N_{WB} \times p_k^{(l)*} \ . \tag{3.5}$$

The released statistics are the averages taken over the implicates

$$B_k^* \equiv \bar{c}_k* = \frac{1}{L}\sum_{l=l}^{L} c_k^{(l)*} \quad \text{and} \tag{3.6}$$

$$\frac{B_k^*}{N_{WB}} \equiv \bar{p}_k* = \frac{1}{L}\sum_{l=l}^{L} p_k^{(l)*} \ . \tag{3.7}$$

Exactly comparable formulas are used for F_k^* and M_k^*.

For each implicate, the finite-population-corrected, *ex-post*-design-weighted sampling variance of the proportion is estimated by assuming that the counts are sampled from a multinomial population and that the missing infrastructure records (equivalent of non-response or coverage errors) are corrected by the QWI weights. Only fuzzed inputs are used in the calculation. Hence, the estimator for the within-implicate variance of the proportion is

$$vp_k^{(l)*} = \left(\frac{p_k^{(l)*}\left(1 - p_k^{(l)*}\right)}{N_{UB}} \right) \left(\frac{N_{WB} - N_{UB}}{N_{WB} - 1} \right) . \tag{3.8}$$

For each implicate, the finite-population-corrected, *ex-post*-design-weighted sampling variance of the count is estimated with

$$vc_k^{(l)*} = N_{WB}^2 \left(\frac{p_k^{(l)*}\left(1 - p_k^{(l)*}\right)}{N_{UB}} \right) \left(\frac{N_{WB} - N_{UB}}{N_{WB} - 1} \right) . \tag{3.9}$$

Again, only fuzzed inputs are used. Notice that the finite population correction (the last term) is not at the cell level. Due to problems with an unknown population count of employment flows within the quarter, we use the state level population correction for all cells. This implicitly assumes that the ratio of the sample to the population is the same as beginning-of-quarter employment, where the population is known.

The Rubin between-variances for the proportions and counts are

$$bp_k^* = \frac{1}{L-1} \sum_{l=1}^{L} \left(p_k^{(l)*} - \bar{p}_k^* \right)^2 \quad \text{and} \tag{3.10}$$

$$bc_k^* = \frac{1}{L-1} \sum_{l=1}^{L} \left(c_k^{(l)*} - \bar{c}_k^* \right)^2 . \tag{3.11}$$

The Rubin average within-variances for the proportions and counts are

$$\bar{vp}_k^* = \frac{1}{L} \sum_{l=1}^{L} vp_k^{(l)*} \quad \text{and} \tag{3.12}$$

$$\bar{vc}_k^* = \frac{1}{L} \sum_{l=1}^{L} vc_k^{(l)*} . \tag{3.13}$$

The Rubin total variances are

$$tvp_k^* = \bar{vp}_k^* + \left(\frac{L+1}{L}\right) bp_k^* \quad \text{and} \tag{3.14}$$

$$tvc_k^* = \bar{vc}_k^* + \left(\frac{L+1}{L}\right) bc_k^* \,. \tag{3.15}$$

For completeness, we also calculate the Rubin missingness ratio as

$$mrp_k^* = \frac{\left(\frac{L+1}{L}\right) bp_k^*}{tvp_k^*} \,, \tag{3.16}$$

and similarly for mrc_k^*.

All formulas for full-quarter employment and total employment, F and M, are comparable – substituting $f_{i,j}^{(l)}$ for $b_{i,j}^{(l)}$, $F_k^{(l)}$ for $B_k^{(l)}$, N_{WF} for N_{WB}, and N_{UF} for N_{UB} in the case of F, with analogous substitutions for M. Because N_{WF} and N_{WM} are not benchmarked by the QCEW but are based on the weights for beginning of-of-quarter employment, there may be negative finite population corrections that we replaced with the smallest positive finite-population correction factor based on B.[8]

3.3.3 Total Variability Model for Z_W3

The cells for $Z\_W3_k^*$ are the same mutually-exclusive and exhaustive cells as used for F_k^*. For any implicate l, the fuzzed-weighted estimate of average monthly earnings is calculated as

$$Z\_W3_k^{(l)*} = \frac{1}{F_k^{(l)}} \sum_{(i,j)\in\{\text{def } k\}} z\_w3_{i,j}^{(l)} w_j \delta_j, \tag{3.17}$$

where $F_k^{(l)}$ is the unfuzzed-weighted full-quarter employment for cell k. To compute the sampling variance of $Z\_W3_k^{(l)*}$, we use the fuzzed-weighted uncorrected sum of squares,

[8]This procedure is essentially the same as the method used for finite population corrections in the American Community Survey (Starsinic, 2011).

calculated as

$$uss_k^{(l)*} = \sum_{(i,j)\in\{\text{def } k\}} \left(z\_w3_{i,j}^{(l)}\right)^2 w_j\delta_j \ . \tag{3.18}$$

For each implicate, the finite-population-corrected, *ex-post*-design-weighted sampling variance of the average monthly earnings for full-quarter employed workers is estimated with

$$vz_k^{(l)*} = \frac{1}{F_k^{(l)u}}\left(\frac{uss_k^{(l)*}}{F_k^{(l)}} - \left(Z\_W3_k^{(l)*}\right)^2\right)\left(\frac{N_{WF} - N_{UF}}{N_{WF} - 1}\right) \tag{3.19}$$

where $F_k^{(l)u}$ is the unfuzzed-unweighted count of full-quarter employment in cell k, and $vz_k^{(l)*}$ is only computed when $F_k^{(l)}$ is positive. Notice that the formula for the within-variance for each implicate is a conditional sampling variance, given membership in cell k. In all cases unfuzzed-weighted values are used in the denominator and fuzzed values (weighted or unweighted) are used in the numerator. This is consistent with the approved QWI noise-infusion system and prevents cancellation of the fuzz-factor when only one establishment populates the cell. Because the average, $Z\_W3_k^{(l)*}$, is computed according to equation 3.17 and the mean uncorrected sum of squares is computed using the same denominator as $Z\_W3_k^{(l)*}$, the term $\left(\frac{uss_k^{(l)*}}{F_k^{(l)}} - \left(Z\_W3_k^{(l)*}\right)^2\right)$ in equation 3.19 can be negative. This situation arises for small values, generally less than three, of $F_k^{(l)}$ when the discrepancy between the fuzzed count $F_k^{(l)*}$ and the unfuzzed count $F_k^{(l)}$ is relatively large. When this happens, the term $\left(\frac{uss_k^{(l)*}}{F_k^{(l)}} - \left(Z\_W3_k^{(l)*}\right)^2\right)$ is set to zero attributing all variation to the between-implicate variance.

The quantities for the Rubin total variance can now be computed for $Z\_W3_k^{(l)*}$. The publication statistic is

$$Z\_W3_k^* \equiv z\_\bar{w}3_k^* = \frac{1}{L}\sum_{l=1}^{L} Z\_W3_k^{(l)*} \ . \tag{3.20}$$

The between-implicate variance is

$$bz_k^* = \frac{1}{L-1}\sum_{l=1}^{L}\left(Z\_W3_k^{(l)*} - z\_\bar{w}3_k^*\right)^2 \ . \tag{3.21}$$

The average within-implicate variance is

$$\bar{vz}_k^* = \frac{1}{L}\sum_{l=1}^{L} vz_k^{(l)*} \ . \tag{3.22}$$

Finally, the Rubin total variance is

$$tvz_k^* = \bar{vz}_k^* + \frac{L+1}{L} bz_k^* \tag{3.23}$$

We also calculate the Rubin missingness ratio for average monthly earnings of full-quarter employees using the formula equivalent to equation 3.16.

3.3.4 Total Variability Model for $W1$ (Total Payroll)

The cells for total payroll, $W1_k^*$, are the same mutually-exclusive and exhaustive cells as used for M_k^*. For any implicate l, the fuzzed-weighted estimate of total payroll is

$$W1_k^{(l)*} = \sum_{(i,j)\in\{\text{def } k\}} w1_{i,j}^{(l)} w_j \delta_j \tag{3.24}$$

where $w1_{i,j}^{(l)}$ is the gross payroll in cell k. To compute the sampling variance of $W1_k^{(l)*}$, we use the average payroll per worker multiplied by an estimate of the number of workers in cell k, $W1_k^{(l)*} = M_k^{(l)*} \times Z\_W1_k^{(l)*}$. First, we require the fuzzed-weighted estimate of average quarterly earnings, which is calculated as

$$Z\_W1_k^{(l)*} = \frac{1}{M_k^{(l)}} \sum_{(i,j)\in\{\text{def } k\}} w1_{i,j}^{(l)} w_j \delta_j \tag{3.25}$$

where $M_k^{(l)}$ is the unfuzzed-weighted employment flow for cell k. We also have the fuzzed-weighted uncorrected sum of squares, calculated as,

$$mss_k^{(l)*} = \sum_{(i,j)\in\{\text{def } k\}} \left(w1_{i,j}^{(l)}\right)^2 w_j \delta_j \tag{3.26}$$

For each implicate, the finite-population-corrected, *ex-post*-design-weighted sampling variance of total payroll is estimated with

$$vw_k^{(l)*} = \frac{\left(M_k^{(l)*}\right)^2}{M_k^{(l)u}} \left(\frac{mss_k^{(l)*}}{M_k^{(l)}} - \left(Z\_W1_k^{(l)*}\right)^2\right)\left(\frac{N_{WM} - N_{UM}}{N_{WM} - 1}\right) \tag{3.27}$$

where N_{WM} and N_{UM} are the fuzzed-weighted count and the fuzzed-unweighted counts of population flows, respectively. The denominator in the first term, $M_k^{(l)u}$, is the unfuzzed-unweighted cell count. The numerator of the first term scales the sample mean to give the sample variance of a count. Just as with $Z\_W3_k^*$, the middle term in 3.27 may be negative, which we then set to zero and attribute all variance to between-implicate variance.

The quantities for the Rubin total variance can now be computed for $W1_k^*$. The publication statistic is

$$W1_k^* \equiv \bar{W1}_k^* = \frac{1}{L}\sum_{l=1}^{L} W1_k^{(l)*} \ . \tag{3.28}$$

The between-implicate variance is

$$bw_k^* = \frac{1}{L-1}\sum_{l=1}^{L}\left(W1_k^{(l)*} - \bar{W1}_k^*\right)^2 \ . \tag{3.29}$$

The average within-implicate variance is

$$\bar{vw}_k^* = \frac{1}{L}\sum_{l=1}^{L} vw_k^{(l)*} \ . \tag{3.30}$$

Just as in equation 3.23, the Rubin total variance is

$$tvw_k^* = \bar{vw}_k^* + \frac{L+1}{L} bw_k^* \ . \tag{3.31}$$

We also calculate the Rubin missingness ratio for average monthly earnings of full-quarter employees using the formula equivalent to equation 3.16.

3.3.5 Reconciling Total Variability Measures Using Published Values of B, F, M, Z_W3, and $W1$

Once we compute the five QWI statistics, we perform quality checks and modify the within- and between-variance so they are consistent with public-use values. For reasons previously discussed, we compute the final total variability statistics using a research

process distinct from the production process used to compute the QWI public-use statistics.[9] The resulting QWI statistics differ in some circumstances from the official public-use statistics, with the most discord occurring in the smallest public-use cells. To scale the internally calculated total variability statistics to the publicly released statistics, we assume the coefficient of variation is equal in both the public-use and internally calculated total variability statistics. In order ensure the reasonableness of this assumption, we edit the coefficient of variation of the QWI statistic when it deviates substantially from similar cells within the same aggregation level, and with the same size QWI statistic.

For each table, we merge a public-use table of QWI statistics with our corresponding internal calculations of the five QWI statistics and their associated total variability measures. Next, we bin each internally calculated employment measure, respectively, by aggregation level and into centiles of employment. We calculate the 5th and 95th percentiles of the coefficient of variation for each bin. Within each bin, we consider cells below the 5th percentile and above the 95th percentile of the coefficient of variation outliers, and we replace their within- and between-variance with the within- and between-variance of the median of coefficient of variation of the bin. We also replace the internal statistic with the value of the corresponding median of the coefficient of variation of the bin. Note that the public-use statistic is always preserved and is the reference statistic for all total variability measures. Appendix C.2 provides a more detailed summary of the procedure.

Before computing the released total variability measures consistent with the public-use QWI, we account for, and flag, the presence of sampling zeros. The public-use QWI contains only cells where at least one statistic is computable for the given cell, which means there is at least one UI-covered job in that cell. The frame for the QWI, however, is establishments whether they have positive UI-covered jobs in a quarter or not. Thus,

[9]To recap, research computing uses a snapshot of a single collection of vintages of the LEHD infrastructure file system that were used to compute one release of the data, in this case R2012Q4. Some production system edits are not captured in this snapshot. Similarly, some research system edits are not reflected in the production system.

it is possible that a given cell will have no released QWI statistics, but nonetheless be at risk for positive employment. This is a sampling zero. In contrast, some cells will never have positive employment or observed firm activity, and we denote these structural zeros because they are not at risk to have any employment in the cell. We flag these two types of cells for advanced users and impute variance measures for the sampling zeros. Appendix C.3 gives a detailed summary of the procedure.

After checks for the quality of the final statistics, we create the released statistics using the edited data and their corresponding statistics when necessary. We only release total variability statistics for unsuppressed statistics in the public-use data. For each total variability statistic, we scale the within- and between-variance by the square of the ratio of the public-use statistic to the internally computed statistic. As long as the public-use value is close to the value we calculate, otherwise we use a representative value from another bin. This follows from our previously stated assumption of equal coefficients of variation within a cell. The total variance, missingness ratio, and degrees of freedom are recalculated from the scaled within- and between-variance. The final file contains the same identifiers, QWI statistics, and status flags as the public-use tables. In addition, it includes the five total variability statistics rounded to three significant digits whenever the public-use statistic is present. The only additional records in the total variability files beyond what is in the public-use QWI are the sampling zeros, which receive variability measures as described in Appendix C.3. The original, unscaled total variability statistics will be used whenever either the public-use or internally calculated statistic are zero.

3.4 Results

We summarize the results in Table 3.1 for all total employment, $EmpTotal$, Table 3.2 for all beginning-of-quarter employment, Emp, in Table 3.3 for all full-quarter employment,

EmpS, Table 3.4 for all total payroll, *Payroll*, in Table 3.5 for all average monthly earnings of full-quarter employees, *EarnS*. Tables showing the same statistics for only private establishments are shown in Appendix Tables C.6 to C.10. In addition to summaries of the statistics defined above, we also summarize the distribution of the coefficient of total variation, which is the square root of the total variance divided by the estimated statistic for *EmpTotal*, *Emp*, *EmpS*, *EarnS*, and *Payroll*. For *Emp* this formula is

$$cvc_k^* = \frac{\sqrt{tvc_k^*}}{Emp_k^*} \tag{3.32}$$

The same equation holds for the four other statistics using their respective total variances in the numerator and the corresponding statistic in the denominator.

3.4.1 Interpretation of the Tables

Tables 3.1-3.5 have the same structure.[10] The major row label is the level of QWI tabulation. For example, the row labelled "Age x Gender" refers to the collection of tabulations stratified by year, quarter, ownership (private), state, age category, and gender. The published QWI data conform to the schema listed here: `http://lehd.ces.census.gov/doc/QWIPU_Data_Schema.pdf`. Refer to this page for categories of the stratifying variables. The minor row label characterizes the publication cell by its size. For Table 3.2 the size classes are based on the values of beginning-of-quarter employment. For Tables 3.1 and 3.4 the size classes are based total employment, and for Tables 3.1 and 3.5, the classes are based on full-quarter employment. The complete set of size classes we summarize is:

- Zero measured value, after rounding, which means that the estimated value is zero.
- 1-2, which means that the estimated value is in the interval [1,2] after rounding.
- 3-9, which means that the estimated value is in the interval [3,9] after rounding.

[10] Appendix Tables C.6 to C.10 also follow this structure.

- 10-99, which means that the estimated value is in the interval [10,99] after rounding.
- 100-999, which means that the estimated value is in the interval [100,999] after rounding.
- +1000, which means that the estimated value is in the interval [1000,max] after rounding.

The column labeled "Proportion of Cells" shows the proportion of all cells in the major row category that lie within the minor row category size class. For example, the value 1.000 in Table 3.1, for the Age x Gender publication tables in the +1000 size class indicates that all the cells in the year x quarter x ownership (all) x state x age category x gender publication tables have at least 1,000 employees in the cell for the publication period 1990:1 through 2012:1. The column labeled "Number of Cells" gives the total number of cells published for this major row category in the indicated count range. Using the same row as an example, the value 46,480 means that there are this many unique cells in the year x quarter x ownership (all) x state x age category x gender publication tables for the same period.

For Tables 3.1, 3.2, and 3.3 the next column is "Median Count," which is the median value of the tabulation variable $EmpTotal$, (respectively, Emp, $EmpS$) in the cells covered by that row. Using the same example row in Table 3.1, the value 91,515 is the median value of total employment in the 46,480 age x gender cells summarized in that row. For Table 3.5, the next column is "Median Average Monthly Earnings," which is the median value of average monthly earnings for all of cells tabulated in a row of the table. For Table 3.4, the next column is "Median Payroll." For all five tables, we report medians rather than averages for most statistics. We compute all tabulations over all tabulated cells used for that row. Upon careful review of the summary tables, we found outlier cells to have undo influence on summary statistics based on averages. We therefore use medians, which believe best summarizes the "typical" cell for a given stratification.

For Tables 3.1-3.5, the next column "Median Total Variation" reports the median value of the Rubin total variation for the cells tabulated in that row. In Tables 3.1, 3.2, and 3.3 this is the median value of tvc_k^* from equation 3.15 variable $EmpTotal$ (respectively, Emp, $EmpS$). In Table 3.4 this is the median value of tvw_k^* from equation 3.31, and from Table 3.5 it is the median value of tvz_k^* from equation 3.23. The values tabulated in this column are the overall summary measures of data quality for the five released total quality measures.

For Tables 3.1-3.5, the next column "Median Rubin Missingness Rate (Percent)" reports the median value of the missingness ratio stated as a percentage. The reported statistic is the median value in a cell over all cells used in the indicated row. See subsection 3.4.3 for a discussion of the interpretation of this data quality statistic.

Again for Tables 3.1-3.5, the next four columns report the "Quantiles of the Coefficient of Variation, where the coefficient of variation is defined in equation 3.32. These columns restate the square root of the Rubin total variation statistic as a ratio to the estimated value of the publication statistic. These statistics on the coefficient of variation can be used to assess the proportionate total variation around the published value arising from all sources of error. See the discussion in sub-section 3.4.2.

The final three columns of Tables 3.1-3.5, "Approximate median 90% Confidence Interval Margin of Error" report the Rubin approximate degrees of freedom and the margins of error of the median 90% approximate confidence intervals. The "margin of error" is one-half of the 90% confidence interval width. For $EmpTotal$, Emp, and $EmpS$, we compute the approximate degrees of freedom using the moment-matching formula from Rubin and Schenker (1986)

$$df_k^* = (L-1)\left(1 + \frac{L}{L+1}\frac{\bar{vc}_k^*}{bc_k^*}\right)^2 \tag{3.33}$$

where the appropriate within-variance (equation 3.13) and between-variance (equation 3.11) is used in the numerator and denominator, respectively. To compute the approxi-

mate degrees of freedom for confidence intervals for *EarnS*, we use the within-variance from equation 3.22 and the between variance from equation 3.21 in equation 3.33. In all cases, $L = 10$. The same logic applies to *Payroll* with its corresponding equations. The margin of error for the count is computed by multiplying the square root of the average total variance by the t-statistic value for probability 0.05 with the degrees of freedom indicated in the "df" column. The margin of error for the percent is calculated by multiplying the average coefficient of variation by the same t-statistic, then expressing the result as a percentage.

The engaged reader may notice a seeming anomaly when viewing the summary median degrees of freedom in Tables 3.1-3.5. The median degrees of freedom for the Industry x County, employment sizes 3-9 row, reside at our imposed upper bound and appear curious compared to the other rows. This is especially true compared to the row above. The Industry x County, employment sizes 1-2 row has a much smaller median degrees of freedom, in line with the other rows in the summary tables. Upon further inspection, this is not an error and the apparent anomaly lies with the suppression rules in the QWI public-use tables and the preponderance of multi-unit employers in a given cell. In the case of the latter, recall that there is no multiple imputation of county or industry in the QWI. The only source of between variance at the Industry x County level is through the imputation of a workplace to a worker – called the unit-to-worker impute in the technical documentation. Cells with employment in the range 3-9, have few firms, and the distribution of firms skews towards single-establishment firms. These firms have no unit-to-worker impute, and are not a source of between variance generating the upper bound on the degrees of freedom. The other important aspect is the suppression of most cells in the public-use data that contain employment counts of 1-2. In the cell counts in Tables 3.1-3.5 one sees a sharp dip in the cell count. This is not a representative sub-population of cells, which leads to anomalous looking summary results. When one looks at Table 3.4, *Payroll*, for which items are never suppressed, one sees that the median degrees of

freedom is also at the upper bound, which is what we would expect given their small employment size.

We interpret the approximate median 90% confidence interval margins of error for the counts as providing evidence about the overall reliability of counts of *EmpTotal*, *Emp*, and *EmpS* for cells that lie in the indicated count range. For example, the margin of error for the count associated with the Age x Gender cell in Table 3.2, +1000 row is 94, and the average value of *Emp* in that row is 70,233. The approximate 90% confidence intervals are 70,233 +/- 94. The approximate confidence interval margins of error for counts are most useful for assessing the reliability of estimates in the range zero (after rounding) to nine, although we provide them for all count ranges.

We interpret the approximate average 90% confidence intervals stated in percentages as providing evidence on the relative reliability of counts of *EmpTotal*, *Emp*, and *EmpS*. Using the same row as an example, we have the relative 90% confidence interval of 70,233 +/- 0.13%. The approximate confidence interval margins of error stated in percentages are useful for assessing the reliability of estimates in the range 10 to 1,000 and over – that is, for the cells containing the vast bulk of employment.

3.4.2 Computing Confidence Bounds for Published Estimates of *EmpTotal*, *Emp*, *EmpS*, *Payroll*, and *EarnS*

In this subsection, we explain how to use the distribution files to compute more accurate 90% confidence intervals for published QWI and LODES data.[11] The distribution files contain total variation measures, computed using equation 3.15) for *EmpTotal*, *Emp*, and *EmpS*, and equation 3.23 for *EarnS*, and equation 3.31 for *Payroll*. The components of the

[11] Found here: `http://doi.org/10.3886/E100590V1`.

confidence intervals used to compute the results in Tables 3.1-3.5 can be replaced by the comparable quantities in the distribution files to improve the accuracy of the confidence intervals.

Find the appropriate distribution table (corresponding to a major row label in Tables 3.1-3.5) and the appropriate rows of the distribution file (corresponding to the desired values of the stratifying variables). Take the square root of the total variation measure to form confidence intervals for the reported values of *EmpTotal*, *Emp*, and *EmpS*, *Payroll*, and *EarnS*. Divide the square root of the total variation measure by the level of the published value to form percentage confidence intervals. Derive the within variance using the total variance and the missingness ratio as

$$\bar{vc}_k^* = (1 - mrc_k^*)\, tvc_k^* \ , \tag{3.34}$$

where the appropriate value of the missingness ratio and the total variance should be used for the different statistics, respectively. Derive the between variance using total variance, within variance and the total number of implicates according to the formula

$$bc_k^* = \frac{L}{L+1}\left(tvc_k^* - \bar{vc}_k^*\right) \ , \tag{3.35}$$

where $L = 10$. Finally, compute the approximate degrees of freedom according to equation 3.33.

To form a more accurate confidence interval for the level of the published indicator, multiply the square root of the total variance for that measure by the appropriate value from the t-distribution with the degrees of freedom indicated by equation 3.33 and the desired confidence level. To form a more accurate confidence interval for the percentage variation of the published indicator, divide the margin of error calculated for the level by the value of the published statistic. We recommend using confidence intervals calculated from employment counts for cells with tabulations from zero to nine. We recommend using confidence intervals calculated from the percentage variation in employment for

cells with tabulations of 10 or more. For confidence intervals on average monthly earnings of full-quarter employment, we recommend using percentage variation.

Users of LODES/OTM can use Table 3.2 to estimate approximate confidence intervals for workplace employment counts published in OTM or calculated directly from LODES. Find the major row label in Table 3.2 that most closely approximates the stratification used in the LODES/OTM workplace summary. Generally, that will be one of the tables with detailed "county-level" geographic stratification combined with demographic or firm-level variables. There is no QWI equivalent for the earnings category stratification available in LODES. Once the closest suitable QWI table has been selected, select the row with the count range that corresponds to the employment count for which a confidence interval is desired. For employment counts of zero to nine, use the count margin of error to form an approximate 90% confidence interval. For employment counts of 10 or more, use the percentage margin of error to form an approximate 90% relative confidence interval. If other levels of confidence are required, use the degrees of freedom estimate in the same row to look up the correct t-statistic for the desired confidence level, then compute count margins of error using the square root of the average total variation in the row or compute percentage margins of error using the average coefficient of variation in the row.

3.4.3 Discussion of the Interpretation of Missingness Ratios and Data Quality

The Rubin total variance measure is the appropriate statistic to summarize the total quality of the published indicators for total employment, beginning-quarter employment, full-quarter employment, total payroll, and average monthly earnings of full-quarter employees. It is clear from Tables 3.1-3.5 that total variation declines monotonically, in percentage terms, as the number of jobs in the tabulation value increases. This is hardly surprising,

but careful attention to the magnitudes of these percentage total variations (in the coefficient of variation columns) shows that for even the most detailed tables and for the stratifiers associated with the largest missingness ratios, the tabulations are very reliable when based on job counts of at least 10, and moderately reliable for job counts of three to nine. This conclusion remains valid even if the very pessimistic assessment of total variation (the 95th percentile of the distribution of the coefficient of variation) is used.

The missingness ratio, therefore, is not a measure of total quality. Instead, it is an indicator of which components of the infrastructure used to compute the QWI and LODES can be most improved by investments in data that reduce the amount of edit and imputation required to estimate that component.

Two components stand out in this regard: education in comparison with worker age and gender. Education is imputed for the vast majority (about 87%) of the individuals in the LEHD infrastructure based on a multistage ignorable missing data model. By contrast, worker age and gender are imputed for less than seven percent of the individuals. And race and ethnicity are imputed for about 18% of the individuals. Looking closely at the average coefficients of variation for the Age x Gender x Industry x County table in comparison with the Gender x Education x Industry x County table, we see that for every count range, the Age x Gender table has less total variation than the Gender x Education table. The explanation is that the missingness ratio never falls below 91% for the Gender x Education table, whereas it varies between 41% and 71% for most of the Age x Gender table. The statistics confirm that the quality of the Gender x Education table can only be improved by reducing the contribution from missing data. The analysis also confirms that even with very large missingness ratios, the Gender x Education tabulations have acceptable total variation for tabulations involving at least 10 employees.

3.5 Conclusion

We have conducted the first comprehensive total quality analysis of five major publication variables in the Quarterly Workforce Indicators, namely the two key employment indicators and the most widely used earnings indicator. The beginning-of-quarter employment variable from QWI is also the primary tabulation variable in the LEHD Origin-Destination Employment Statistics; hence, our analysis is also applicable to workplace tabulations directly from LODES or displayed in OnTheMap. Our analysis reveals that the very smallest tabulations (estimated zeros and counts of one or two) are not particularly reliable in the sense that they could easily range from zero to three. Tabulations of three to nine are more reliable in the sense that the 90% confidence bound is generally less than plus or minus four. Tabulations involving 10 or more jobs are very reliable having percentage variation that declines from a worst case of plus or minus 31% (count range 10-99, tables involving education) to a best case of plus or minus less than one percent (count range +1000, tables involving firm age).

To the best of our knowledge, no other widely used statistical system based on administrative records has produced a comprehensive total variation analysis to which the results in this paper can be compared. As compared to survey-based estimates like those derived from the American Community Survey, for example, the QWI employment and earnings tabulations have accuracy comparable to the PUMA and small state accuracy of the ACS (U.S. Census Bureau, 2015) even though the QWI total variability measures include the errors from coverage, edit, imputation, and sampling while those from the ACS include only sampling variability.

3.6 Summary Tables

Table 3.1: Summary of Total Variability of All Total Employment (*EmpTotal*) by Table and Count

Table and *EmpTotal* count range	Proportion of Cells	Number of Cells	Median Count	Median Total Variation	Median Rubin Missingness Rate (Percent)	Quantiles of Coefficient of Variation			Median Approximate 90% Confidence Intervals Margin of Error		
						5th	Median	95th	Median df	Count	Percent
All (Private plus State and Local)											
Age x Gender											
+1000	1.0000	46,480	91,515	8690.00	43.10%	0.0003	0.0010	0.0032	48	121	0.13%
Race x Ethnicity											
10-99	0.0181	632	56	51.55	96.70%	0.0837	0.1403	0.2568	9	10	19.40%
100-999	0.1223	4,263	443	415.00	95.60%	0.0265	0.0474	0.0932	9	28	6.56%
+1000	0.8596	29,965	14,956	6310.00	87.30%	0.0002	0.0041	0.0269	11	108	0.56%
Gender x Education											
+1000	1.0000	23,240	187,994	222000.00	96.80%	0.0012	0.0028	0.0079	9	652	0.39%
Industry x County											
zero measured value, after rounding	0.0026	8,225	0	0.31	94.90%	(a)	(a)	(a)	10	1	(a)
1-2	0.0000	134	1	0.39	80.10%	0.1664	0.3652	0.9434	14	1	49.21%
3-9	0.0132	41,946	7	0.51	0.00%	0.0593	0.1075	0.3871	9999	1	13.78%
10-99	0.2333	743,122	47	5.52	43.50%	0.0237	0.0537	0.1643	47	3	6.98%
100-999	0.4539	1,445,825	307	57.90	70.80%	0.0101	0.0238	0.0593	18	10	3.17%
+1000	0.2971	946,236	2,989	774.00	77.10%	0.0023	0.0080	0.0199	15	37	1.08%
Age x Gender x Industry x County											
zero measured value, after rounding	0.1672	7,973,123	0	0.21	95.20%	(a)	(a)	(a)	9	1	(a)
1-2	0.0049	234,460	2	0.38	72.70%	0.1527	0.3252	0.7382	17	1	43.36%
3-9	0.2165	10,324,414	5	0.85	66.20%	0.0864	0.1754	0.3944	20	1	23.24%
10-99	0.4014	19,140,564	27	5.33	71.40%	0.0367	0.0806	0.1803	17	3	10.75%
100-999	0.1737	8,279,653	224	52.30	75.60%	0.0137	0.0294	0.0610	15	10	3.95%
+1000	0.0363	1,728,489	1,982	482.00	76.00%	0.0041	0.0101	0.0197	15	29	1.35%
Race x Ethnicity x Industry x County											
zero measured value, after rounding	0.5635	19,553,448	0	0.20	95.50%	(a)	(a)	(a)	9	1	(a)
1-2	0.0062	216,005	2	0.70	92.10%	0.2688	0.6245	0.9354	10	1	85.69%
3-9	0.1400	4,856,222	5	2.34	89.50%	0.1334	0.3159	0.5944	11	2	43.07%
10-99	0.1653	5,735,093	26	10.10	86.20%	0.0431	0.1169	0.2692	12	4	15.85%
100-999	0.0886	3,073,969	254	75.20	80.50%	0.0132	0.0317	0.0736	13	12	4.29%
+1000	0.0364	1,262,815	2,573	745.00	79.60%	0.0031	0.0093	0.0210	14	37	1.25%
Gender x Education x Industry x County											
zero measured value, after rounding	0.0737	1,787,333	0	0.26	95.10%	(a)	(a)	(a)	9	1	(a)
1-2	0.0044	106,593	2	1.38	93.40%	0.4290	0.6538	0.9513	10	2	89.72%
3-9	0.1901	4,610,815	5	4.05	93.20%	0.2386	0.3783	0.6101	10	3	51.91%
10-99	0.4433	10,755,591	29	22.50	93.20%	0.0853	0.1597	0.2946	10	7	21.91%
100-999	0.2305	5,593,317	234	187.00	93.40%	0.0291	0.0566	0.0963	10	19	7.76%
+1000	0.0580	1,407,901	2,090	1770.00	93.70%	0.0084	0.0192	0.0318	10	58	2.63%

Notes: Total employment is defined as all jobs held by a worker at the same establishment during the quarter. Statistics are computed across all state-year-quarters within a table. The "All" category of establishments includes private as well as state and local government but excludes federal employment. All tables include all valid QWI age groups with the exception of any table including education, in which case only jobs with workers age 25 and older are included. For statistic definitions for total employment, please see their respective equations in the accompanying text: Count 3.6, Total Variation 3.15, Missingness Ratio 3.16, Coefficient of Variation 3.32. (a) Undefined value.

Table 3.2: Summary of Total Variability of All Beginning-of-Quarter Employment (*Emp*) by Table and Count

Table and *Emp* count range	Proportion of Cells	Number of Cells	Median Count	Median Total Variation	Median Rubin Missingness Rate (Percent)	Quantiles of Coefficient of Variation			Median Approximate 90% Confidence Intervals Margin of Error		
						5th	Median	95th	Median df	Count	Percent
All (Private plus State and Local)											
Age x Gender											
+1000	1.0000	45,712	70,233	5300.00	37.00%	0.0003	0.0010	0.0032	65	94	0.13%
Race x Ethnicity											
10-99	0.0258	883	51	39.50	96.50%	0.0793	0.1277	0.2664	9	9	17.66%
100-999	0.1489	5,105	454	326.00	95.10%	0.0198	0.0430	0.0830	9	25	5.95%
+1000	0.8253	28,296	12,858	4340.00	84.60%	0.0001	0.0038	0.0237	12	89	0.52%
Gender x Education											
+1000	1.0000	22,856	161,812	162000.00	96.80%	0.0012	0.0028	0.0079	9	557	0.39%
Industry x County											
zero measured value, after rounding	0.0056	17,598	0	0.28	95.50%	(a)	(a)	(a)	9	1	(a)
1-2	0.0001	257	2	0.44	78.30%	0.1443	0.3592	0.8972	14	1	48.31%
3-9	0.0203	63,664	7	0.43	0.00%	0.0546	0.1022	0.3814	9999	1	13.09%
10-99	0.2633	827,121	45	5.06	50.30%	0.0223	0.0529	0.1643	35	3	6.91%
100-999	0.4464	1,402,205	295	55.70	74.70%	0.0099	0.0240	0.0590	16	10	3.20%
+1000	0.2643	830,357	2,875	711.00	79.30%	0.0023	0.0080	0.0197	14	36	1.07%
Age x Gender x Industry x County											
zero measured value, after rounding	0.2011	9,317,087	0	0.20	95.80%	(a)	(a)	(a)	9	1	(a)
1-2	0.0051	234,090	2	0.36	74.20%	0.1371	0.3156	0.7273	16	1	42.19%
3-9	0.2246	10,406,647	5	0.77	67.90%	0.0794	0.1675	0.3873	19	1	22.24%
10-99	0.3842	17,797,008	27	4.99	74.40%	0.0351	0.0791	0.1793	16	3	10.58%
100-999	0.1547	7,165,326	222	48.90	77.90%	0.0131	0.0288	0.0603	14	9	3.87%
+1000	0.0303	1,405,442	1,945	437.00	77.60%	0.0039	0.0097	0.0192	14	28	1.31%
Race x Ethnicity x Industry x County											
zero measured value, after rounding	0.6023	20,718,981	0	0.19	96.00%	(a)	(a)	(a)	9	1	(a)
1-2	0.0056	191,678	2	0.67	92.70%	0.2632	0.6050	0.9028	10	1	83.01%
3-9	0.1288	4,431,864	5	2.16	90.10%	0.1256	0.3044	0.5799	11	2	41.50%
10-99	0.1514	5,208,590	26	9.27	86.80%	0.0402	0.1115	0.2610	11	4	15.20%
100-999	0.0805	2,767,906	251	69.40	82.00%	0.0126	0.0307	0.0710	13	11	4.15%
+1000	0.0314	1,081,496	2,506	673.00	81.30%	0.0030	0.0091	0.0204	13	35	1.23%
Gender x Education x Industry x County											
zero measured value, after rounding	0.0870	2,055,422	0	0.26	95.70%	(a)	(a)	(a)	9	1	(a)
1-2	0.0049	114,711	2	1.37	94.20%	0.4269	0.6496	0.9421	10	2	89.14%
3-9	0.2033	4,805,343	5	3.93	93.90%	0.2359	0.3758	0.6065	10	3	51.56%
10-99	0.4392	10,378,260	29	21.50	94.00%	0.0846	0.1597	0.2935	10	6	21.91%
100-999	0.2146	5,070,981	231	180.00	94.10%	0.0286	0.0561	0.0953	10	18	7.69%
+1000	0.0511	1,207,342	2,051	1660.00	94.40%	0.0085	0.0190	0.0313	10	56	2.60%

Notes: Beginning-of-quarter employment is defined as all jobs held by a worker at the same establishment during the quarter and during the previous quarter. Statistics are computed across all state-year-quarters within a table. The "All" category of establishments includes private as well as state and local government but excludes federal employment. All tables include all valid QWI age groups with the exception of any table including education, in which case only jobs with workers age 25 and older are included. For statistic definitions for beginning of quarter employment, please see their respective equations in the accompanying text: Count 3.6, Total Variation 3.15, Missingness Ratio 3.16, Coefficient of Variation 3.32. (a) Undefined value.

Table 3.3: Summary of Total Variability of All Full-Quarter Employment (*EmpS*) by Table and Count

Table and *EmpS* count range	Proportion of Cells	Number of Cells	Median Count	Median Total Variation	Median Rubin Missingness Rate (Percent)	Quantiles of Coefficient of Variation			Median Approximate 90% Confidence Intervals Margin of Error		
						5th	Median	95th	Median df	Count	Percent
All (Private plus State and Local)											
Age x Gender											
100-999	0.0001	3	961	402.00	79.30%	0.0209	0.0209	0.0209	14	27	2.81%
+1000	0.9999	44,941	56,533	4060.00	32.10%	0.0003	0.0011	0.0035	87	82	0.14%
Race x Ethnicity											
zero measured value, after rounding	0.0002	7	9	5.16	95.20%	0.2022	0.2589	0.4127	9	3	35.81%
10-99	0.0323	1,088	48	35.15	95.70%	0.0737	0.1267	0.2891	9	8	17.53%
100-999	0.1687	5,685	455	299.00	94.60%	0.0122	0.0420	0.0848	10	24	5.76%
+1000	0.7989	26,928	11,454	3550.00	81.90%	0.0001	0.0039	0.0235	13	80	0.52%
Gender x Education											
+1000	1.0000	22,472	143,578	134000.00	96.60%	0.0012	0.0029	0.0081	9	506	0.39%
Industry x County											
zero measured value, after rounding	0.0085	26,395	0	0.27	95.50%	(a)	(a)	(a)	9	1	(a)
1-2	0.0001	445	2	0.20	0.00%	0.1178	0.2565	0.8139	9999	1	32.88%
3-9	0.0273	84,593	7	0.41	0.00%	0.0557	0.1030	0.3814	9999	1	13.20%
10-99	0.2858	884,129	44	5.06	51.60%	0.0228	0.0539	0.1654	33	3	7.05%
100-999	0.4368	1,351,160	287	55.60	75.30%	0.0101	0.0245	0.0595	15	10	3.29%
+1000	0.2414	746,888	2,812	690.00	79.00%	0.0024	0.0081	0.0198	14	35	1.08%
Age x Gender x Industry x County											
zero measured value, after rounding	0.2313	10,443,994	0	0.20	95.80%	(a)	(a)	(a)	9	1	(a)
1-2	0.0052	234,881	2	0.36	74.00%	0.1383	0.3166	0.7228	16	1	42.33%
3-9	0.2281	10,301,188	5	0.75	67.40%	0.0789	0.1672	0.3849	19	1	22.20%
10-99	0.3679	16,614,512	26	4.96	74.30%	0.0352	0.0799	0.1806	16	3	10.68%
100-999	0.1410	6,367,152	220	48.10	77.50%	0.0130	0.0288	0.0605	15	9	3.86%
+1000	0.0265	1,197,767	1,914	417.00	76.90%	0.0040	0.0097	0.0191	15	27	1.30%
Race x Ethnicity x Industry x County											
zero measured value, after rounding	0.6294	21,431,041	0	0.19	96.00%	(a)	(a)	(a)	9	1	(a)
1-2	0.0053	179,385	2	0.66	92.60%	0.2632	0.6042	0.8922	10	1	82.90%
3-9	0.1210	4,119,278	5	2.08	89.70%	0.1232	0.2998	0.5754	11	2	40.88%
10-99	0.1417	4,825,719	26	8.94	86.00%	0.0393	0.1088	0.2576	12	4	14.76%
100-999	0.0746	2,541,058	249	67.80	81.50%	0.0126	0.0305	0.0701	13	11	4.11%
+1000	0.0281	955,663	2,460	637.00	80.80%	0.0031	0.0091	0.0202	13	34	1.22%
Gender x Education x Industry x County											
zero measured value, after rounding	0.0989	2,281,075	0	0.26	95.60%	(a)	(a)	(a)	9	1	(a)
1-2	0.0053	121,268	2	1.37	94.10%	0.4295	0.6500	0.9421	10	2	89.19%
3-9	0.2129	4,907,073	5	3.90	93.90%	0.2359	0.3763	0.6074	10	3	51.64%
10-99	0.4341	10,007,726	28	21.10	93.90%	0.0849	0.1608	0.2945	10	6	22.06%
100-999	0.2026	4,671,711	229	178.00	94.10%	0.0286	0.0562	0.0953	10	18	7.71%
+1000	0.0462	1,065,581	2,021	1620.00	94.30%	0.0087	0.0190	0.0312	10	55	2.60%

Notes: Total employment is defined as all jobs held by a worker at the same establishment during the quarter. Statistics are computed across all state-year-quarters within a table. The "All" category of establishments includes private as well as state and local government but excludes federal employment. All tables include all valid QWI age groups with the exception of any table including education, in which case only jobs with workers age 25 and older are included. For statistic definitions for total employment, please see their respective equations in the accompanying text: Count 3.6, Total Variation 3.15, Missingness Ratio 3.16, Coefficient of Variation 3.32. (a) Undefined value.

Table 3.4: Summary of Total Variability of All Total Payroll (*Payroll*) by Table and Count

Table and *EmpTotal* count range	Proportion of Cells	Number of Cells	Median Payroll	Median Total Variation	Median Rubin Missingness Rate (Percent)	Quantiles of Coefficient of Variation			Median Approximate 90% Confidence Intervals Margin of Error		
						5th	Median	95th	Median df	Count	Percent
All (Private plus State and Local)											
Age x Gender											
+1000	1.0000	46,480	431,844,381.50	4.06E+11	30.00%	0.0004	0.0014	0.0078	99	822,066.74	0.18%
Race x Ethnicity											
10-99	0.0181	632	248,224.00	2.23E+09	97.70%	0.1058	0.2015	0.4318	9	65,310.59	27.87%
100-999	0.1223	4,263	2,153,721.00	1.82E+10	96.30%	0.0348	0.0672	0.1469	9	186,580.78	9.29%
+1000	0.8596	29,965	80,813,010.00	4.83E+11	83.80%	0.0004	0.0063	0.0449	12	942,546.65	0.85%
Gender x Education											
+1000	1.0000	23,240	1,344,933,652.50	2.35E+13	96.60%	0.0016	0.0038	0.0110	9	6,704,480.56	0.53%
Industry x County											
zero measured value, after rounding	0.0024	8,225	0.00	9.59E+06	99.80%	0.0529	0.4142	1.2418	9	4,282.93	57.28%
1-2	0.0886	309,518	47,962.00	1.43E+07	0.00%	0.0000	0.0656	0.5918	9999	4,846.55	8.41%
3-9	0.0120	41,946	27,741.50	6.22E+06	0.00%	0.0299	0.0854	0.5657	9999	3,196.39	10.95%
10-99	0.2126	743,122	223,252.00	1.72E+08	56.10%	0.0193	0.0590	0.2425	28	17,213.63	7.74%
100-999	0.4137	1,445,825	1,652,146.00	2.85E+09	83.00%	0.0100	0.0314	0.0903	13	72,079.42	4.25%
+1000	0.2708	946,236	19,043,488.50	5.83E+10	82.90%	0.0033	0.0111	0.0315	13	326,004.16	1.49%
Age x Gender x Industry x County											
zero measured value, after rounding	0.1426	7,973,123	0.00	5.66E+05	99.90%	0.0000	0.3606	5.6209	9	1,040.49	49.87%
1-2	0.1515	8,471,709	4,432.00	4.34E+05	52.60%	0.0000	0.1250	0.9335	32	862.07	16.35%
3-9	0.1846	10,324,414	17,514.00	1.00E+07	87.70%	0.0458	0.1819	0.6249	11	4,311.55	24.80%
10-99	0.3423	19,140,564	116,677.00	1.23E+08	87.70%	0.0309	0.0949	0.2717	11	15,121.17	12.94%
100-999	0.1481	8,279,653	1,215,134.00	2.00E+09	87.90%	0.0131	0.0368	0.0921	11	60,974.46	5.01%
+1000	0.0309	1,728,489	14,814,716.00	3.50E+10	85.30%	0.0045	0.0126	0.0311	12	253,725.03	1.71%
Race x Ethnicity x Industry x County											
zero measured value, after rounding	0.4663	19,553,448	0.00	3.51E+06	100.00%	0.0387	0.4623	2.8196	9	2,591.10	63.93%
1-2	0.1778	7,456,341	5,093.00	8.21E+06	98.60%	0.0824	0.6106	1.2213	9	3,962.81	84.45%
3-9	0.1158	4,856,222	21,430.50	6.60E+07	97.00%	0.1083	0.4033	0.8508	9	11,235.78	55.77%
10-99	0.1368	5,735,093	129,598.00	3.71E+08	94.20%	0.0441	0.1494	0.3933	10	26,430.12	20.49%
100-999	0.0733	3,073,969	1,398,948.00	3.58E+09	89.90%	0.0141	0.0421	0.1103	11	81,578.26	5.74%
+1000	0.0301	1,262,815	16,919,665.00	5.45E+10	86.00%	0.0039	0.0127	0.0339	12	316,612.12	1.72%
Gender x Education x Industry x County											
zero measured value, after rounding	0.0639	1,787,333	240.00	2.85E+05	99.60%	0.0000	0.2135	1.9833	9	738.34	29.53%
1-2	0.1360	3,803,163	7,036.00	2.13E+07	98.90%	0.3025	0.6337	1.2283	9	6,382.94	87.65%
3-9	0.1649	4,610,815	25,593.00	1.41E+08	98.10%	0.2569	0.4754	0.8347	9	16,422.56	65.75%
10-99	0.3847	10,755,591	165,697.00	1.13E+09	97.50%	0.1005	0.2040	0.4132	9	46,491.16	28.21%
100-999	0.2001	5,593,317	1,541,163.00	1.33E+10	97.30%	0.0356	0.0740	0.1405	9	159,498.65	10.23%
+1000	0.0504	1,407,901	17,357,576.00	2.23E+11	96.80%	0.0112	0.0266	0.0535	9	653,105.94	3.67%

Notes: Total Payroll is defined only over total employment. It is calculated by summing the earnings for the reference quarter for total employment. See the table on total employment for the relevent counts. Statistics are computed across all state-year-quarters within a table. The "All" category of establishments includes private as well as state, and local government but excludes federal employment. All tables include all valid QWI age groups with the exception of any table including education, in which case only jobs with workers age 25 and older are included. For statistic definitions for beginning of quarter employment, please see their respective equations in the accompanying text: Total payroll 3.28, Total Variation 3.31, Missingness Ratio 3.16, Coefficient of Variation 3.32. (a) Undefined value.

Table 3.5: Summary of Total Variability of All Average Monthly Earnings (*EarnS*) by Table and Count

Table and *EmpS* count range	Proportion of Cells	Number of Cells	Median Average Monthly Earnings	Median Total Variation	Median Rubin Missingness Rate (Percent)	Quantiles of Coefficient of Variation			Median Approximate 90% Confidence Interval Margin of Error		
						5th	Median	95th	Median df	Count	Percent
All (Private plus State and Local)											
Age x Gender											
100-999	0.0001	3	1,779.00	15,000.00	87.00%	0.0688	0.0688	0.0688	11	166.99	9.38%
+1000	0.9999	44,941	2,176.00	5.39	23.50%	0.0004	0.0012	0.0062	164	2.99	0.15%
Race x Ethnicity											
3-9	0.0002	7	2,335.00	307,000.00	97.10%	0.1820	0.3357	0.6261	9	766.30	46.42%
10-99	0.0323	1,088	2,127.50	68,000.00	96.20%	0.0616	0.1201	0.3266	9	360.65	16.62%
100-999	0.1687	5,685	2,225.00	8,170.00	94.70%	0.0133	0.0406	0.0979	10	124.03	5.57%
+1000	0.7989	26,928	2,508.50	140.00	75.30%	0.0004	0.0046	0.0302	15	15.86	0.61%
Gender x Education											
+1000	1.0000	22,472	2,844.00	63.60	94.70%	0.0011	0.0028	0.0080	10	10.94	0.38%
Industry x County											
zero measured value, after rounding	0.0013	4,351	0.00	2,490,000.00	99.50%	(a)	(a)	(a)	9	2182.38	(a)
1-2	0.0859	288,667	2,135.00	6,710.00	0.00%	0.0000	0.0520	0.3027	9999	104.98	6.66%
3-9	0.0252	84,593	1,520.00	9,030.00	0.00%	0.0196	0.0665	0.2977	9999	121.79	8.53%
10-99	0.2632	884,129	1,969.00	6,060.00	44.00%	0.0147	0.0405	0.1292	46	101.22	5.26%
100-999	0.4022	1,351,160	2,264.00	1,810.00	71.30%	0.0071	0.0197	0.0531	17	56.73	2.62%
+1000	0.2223	746,888	2,722.00	337.00	71.50%	0.0022	0.0071	0.0201	17	24.48	0.95%
Age x Gender x Industry x County											
zero measured value, after rounding	0.0016	69,616	0.00	2,790,000.00	99.70%	(a)	(a)	(a)	9	2310.11	(a)
1-2	0.2095	9,158,213	1,276.00	9,220.00	12.10%	0.0000	0.0852	0.4918	611	123.19	10.93%
3-9	0.2357	10,301,188	1,469.00	18,700.00	73.50%	0.0298	0.0971	0.3096	16	182.80	12.97%
10-99	0.3801	16,614,512	1,868.00	8,520.00	76.90%	0.0197	0.0532	0.1441	15	123.74	7.13%
100-999	0.1457	6,367,152	2,383.00	2,040.00	77.40%	0.0079	0.0207	0.0526	15	60.55	2.77%
+1000	0.0274	1,197,767	3,152.00	481.00	73.30%	0.0029	0.0075	0.0191	16	29.32	1.01%
Race x Ethnicity x Industry x County											
zero measured value, after rounding	0.0027	51,636	0.00	6,010,000.00	99.90%	(a)	(a)	(a)	9	3390.54	(a)
1-2	0.3472	6,643,546	1,892.00	252,000.00	97.50%	0.0570	0.2777	0.7523	9	694.27	38.40%
3-9	0.2153	4,119,278	2,009.00	132,000.00	93.90%	0.0580	0.1859	0.4701	10	498.54	25.51%
10-99	0.2522	4,825,719	2,145.00	25,100.00	87.90%	0.0248	0.0748	0.2077	11	216.01	10.20%
100-999	0.1328	2,541,058	2,324.00	2,840.00	80.90%	0.0087	0.0238	0.0624	13	71.95	3.21%
+1000	0.0499	955,663	2,763.00	430.00	75.90%	0.0027	0.0079	0.0212	15	27.80	1.06%
Gender x Education x Industry x County											
zero measured value, after rounding	0.0021	53,160	0.00	3,550,000.00	98.80%	(a)	(a)	(a)	9	2605.83	(a)
1-2	0.1652	4,096,117	1,803.00	451,000.00	98.20%	0.1473	0.3915	0.8583	9	928.79	54.15%
3-9	0.1979	4,907,073	1,899.00	237,000.00	96.20%	0.1280	0.2628	0.5388	9	673.29	36.35%
10-99	0.4035	10,007,726	2,205.00	57,100.00	94.70%	0.0499	0.1110	0.2480	10	327.89	15.22%
100-999	0.1884	4,671,711	2,580.00	10,400.00	94.60%	0.0187	0.0411	0.0878	10	139.94	5.63%
+1000	0.0430	1,065,581	3,188.00	2,370.00	94.40%	0.0066	0.0160	0.0374	10	66.80	2.20%

Notes: Average Monthly Earnings is defined only over full-quarter jobs. It is calculated by taking the earnings for the reference quarter for full-quarter jobs and dividing by 3. See the table on full-quarter employment for the relevent counts. Statistics are computed across all state-year-quarters within a table. The "All" category of establishments includes private as well as state, and local government but excludes federal employment. All tables include all valid QWI age groups with the exception of any table including education, in which case only jobs with workers age 25 and older are included. For statistic definitions for beginning of quarter employment, please see their respective equations in the accompanying text: Average Monthly Earnings 3.20, Total Variation 3.23, Missingness Ratio 3.16, Coefficient of Variation 3.32. (a) Undefined value.

APPENDIX A

APPENDIX: HOURS OFF THE CLOCK

A.1 Models of Efficiency Wages and Labor Hoarding

A.1.1 Model of Increased Worker Effort

The model is intentionally simple, and as such does not present a formal theory of unemployment determination, or any equilibrium outcomes. What it does is clarify the continuous implicit bargain between workers and firms, and give intuition for the difference between hours worked and hours paid. The model is similar to the one presented in Lazear et al. (2015), but here effort takes the form of hours worked above what is explicitly recorded by employers. Workers are assumed to be paid a fixed salary implying that their wage is really a gross earnings measure quoted over a fixed period that does not vary with hours worked. Employer-recorded weekly hours are therefore an estimate usually determined by prevailing laws and/or employer knowledge. In the United States, it is typical for a salaried employee to be quoted a salary on a yearly basis (usually paid every two weeks or twice per month), with 40 hours per week that is loosely monitored by the employer. Workers therefore have some latitude to choose the hours they actually work in a given week, trading off their distaste for work against employers' expectations of output.

In the model, time is discrete and workers are already matched with firms for a negotiated period earnings measure, which I assume is fixed. Workers are exempt from overtime and employers assume workers put in at least the statutory overtime limit of $\bar{h}$ hours. Employers would like to terminate employees deemed to be shirking. Employers monitor worker effort and they use observed hours as a proxy, whether perfectly

observed, or a noisy measure. For example, a worker who chooses not to show up to work stands a high risk of being fired. Conversely, a worker who chooses to put in more hours likely produces more, and sends a (possibly) valuable signal to employers about her ability to produce, which reduces her probability of being fired. I define the function $P(h) : \mathbb{R}_+ \to [0, 1] \subset \mathbb{R}$, which maps hours worked to the probability a worker retains her job in a given period. The probability of retention is an increasing and concave function of hours worked with $P(0) = 0$ and $P'(h) > 0, P''(h) < 0$.

Putting forth effort is costly for workers, both because work is generally unpleasant, and because workers are paid a lump sum regardless of how many hours they actually work. Therefore, workers would prefer to work as little hours as possible. I capture the cost to workers of putting forth greater effort by the function $c(h) : \mathbb{R}_+ \to \mathbb{R}_+$. The costs to greater working hours are increasing and convex with $c(0) = 0$ and $c'(h) > 0, c''(h) > 0$. In addition to the direct costs of more hours, workers must also weigh the costs of unemployment. More specifically, the costs of losing one's job in times of greater labor market slack are much greater than when labor markets are tight, and the probability of contacting and finding a suitable job is much higher. Define the unemployment rate by u, which both employers and workers take as given. Further, define the option value of search from unemployment by R. Although this sounds like the beginnings of a rudimentary search model, no formal theory of of the aggregate determination of vacancies, unemployment and search will be presented here.

Workers choose the optimal hours to work conditional on the explicit costs of working and the indirect costs of aggregate labor market conditions. The problem of the worker is

$$\max_h \quad V(h) = P(h)\,[W - c(h)] + (1 - P(h))(1 - u)R\,, \tag{A.1}$$

where W is the gross earnings in the period, and $V(h)$ is the asset value to the worker of holding the job. The worker chooses the optimal hours of work to maximize A.1. Using

the first order condition and the implicit function theorem yields the following equation,

$$\frac{\partial h}{\partial u} = -\frac{P'(h)R}{\frac{\partial^2 V(h)}{\partial h}} > 0 \quad . \tag{A.2}$$

The inequality follows from the fact that $W - c(h) > (1-u)R$ in order for the employment relationship to continue. Higher unemployment leads to greater hours worked.

Intuitively, hours increase in order to the decrease the probability of being fired, due to the decrease in the payoff from unemployment. The second term in equation A.1 gives the value of outside employment, R. As the probability of ascending to employment decreases – as u increases – workers put forth more hours to avoid unemployment. With the employer's report of hours paid fixed at $\bar{h}$, higher unemployment rates lead to a greater spread between hours worked and hours paid.

A.1.2 Model of Labor Hoarding

The labor hoarding model is best conceptualized using a model of labor demand with adjustment costs to employment. I describe the problem and the result loosely through the "gap approach" (Caballero and Engel, 1993; Caballero et al., 1997). Consider a firm whose production technology in a period uses only total labor inputs $f(m_t)$, with $f_m(\cdot) > 0$, and $m_t = e_t h_t$, where e_t is total employment and h_t is average hours worked. I assume there is an upper bound on the average hours worked per worker in a period so that firms seek to employ more than one worker. Further, define $\bar{h}$ as the long-run average hours per employee in the firm. For the purposes of this discussion, we can assume $\bar{h} = 40$ and that these are recorded by the firm as hours paid regardless of actual hours worked. Production faces aggregate shocks A_t drawn from a probability distribution $F(A)$. I assume only that the probability distribution and production function are such that $f_A(\cdot) > 0$, which implies that negative aggregate shocks should lead to decreased production in the absence of adjustment costs.

Consider a firm faced with negative aggregate demand shock in the next period. Suppose the current period's employment as e_{t-1} and without a loss of generality that $h_{t-1} = \bar{h}$. Following the gap approach, define e_t^* as the *frictionless* optimal employment target for next period given the realization of the negative aggregate demand shock. This is the optimal employment for the firm if for the current period, and the current period only, the firm did not have to pay any costs to adjust employment. Given the nature of the production function, it holds that $e_t^* \leq e_{t-1}$, and $e_t \in [e_t^*, e_{t-1}]$. In other words, the optimal employment level given the shock will be bounded above by the current employment level, and bounded below by the optimal frictionless level. The firm would like to get as close to its optimal production level as possible, but in most cases the optimal employment level will lie above the frictionless level. Thus, firms will adjust average hours down in order to further decrease production beyond what was possible by only adjusting labor yielding $h_t \leq h_{t-1}$. The same logic, but in the opposite direction holds for a positive aggregate demand shock.

Putting this altogether, we see that a negative aggregate demand shock leads to less hours worked. Note that in this model employment and hours move in the same direction in response to a negative aggregate demand shock. The empirical results posit a relationship between labor market slack and the difference between hours paid and hours worked. Given the model, we need to assume the unemployment rate is negatively correlated with aggregate demand shocks. That is, a positive demand shock produces non-positive movement in the unemployment rate. This assumption seems straightforward.

To wrap up, a negative aggregate demand shock raises unemployment and forces firms to cut production as well as raising the unemployment rate. Employment cannot fall to its frictionless level so firms reduce hours worked to get closer to the optimal level of output. As long as hours paid stay near constant, this produces the desired empirical test of greater labor market slack and a non-positive gap between hours worked and

hours paid.

A.2 Details on Inverse Probability Weighting

The ACS links to the LEHD via a protected identification key (PIK). A PIK is a random number mapped to a social security number, which serves as an internal Census Bureau person-level identifier. PIK assignment is usually assumed to be correct, however survey responses who fail to receive a PIK are known to be missing at random in the sense of Rubin (1987).

Following (Meyer and Goerge, 2011) who point out that PIKs appear to be missing at random, I use inverse probability weights to correct the ACS weights for PIKs missing at random. I fit a probit model with a rich set of characteristics to predict the probability of receiving a PIK. Using the fit model, I estimate the predicted probability for each cell, and then multiply the ACS weights by the inverse of the probability of receiving a PIK for each cell. Results obtained using weights adjusted in this manner differ very little from when they are omitted. At no point do the qualitative findings change.

The terms of my data use for this project preclude the disclosure of ACS estimates without commingled LEHD data. As such, I am not able to disclose the probit estimates for the inverse probability weighting. The variables are summarized in Table A.4. However, I am able to describe the key parameters qualitatively. In general, the included covariates and results adhere closely to those used in Meyer and Goerge (2011).

The probability of receiving a PIK increases with age, likely due to greater work experience. The same holds true with education. The probability of receiving a PIK increases with the highest level of educational attainment. People of color and Hispanics are slightly less likely to receive a PIK than white non-Hispanics. Women and American

citizens are more likely to receive a PIK, while respondents who report speaking English "Well" or "Very Well" are more likely to receive a PIK than respondents who report speaking English "Not Well" or "Not at All". Finally, respondents who did not move residences in the last year have a higher probability of receiving a PIK than those who did.

A.3 Details on Hourly/Nonhourly Imputation

Neither the ACS nor the LEHD datasets provide information on frequency or method of pay for their earnings variables. This section describes the imputation of the probability that an ACS respondent was not paid by the hour. This can include earnings or salaries paid at annual, monthly, weekly or biweekly rate. This would also include workers who are paid a piece rate. I use the Current Population Survey Outgoing Rotation Group (CPS ORG) files, which asks respondents if they were paid by the hour or by some other arrangement on their main job last week. I model the probability that a job is not paid by the hour using a logit model, with covariates describing firm and job characteristics common to both the CPS ORG and ACS. After fitting the model, I generate a predicted probability for each cell of covariates. The final estimates are then attached to the ACS using the common covariates. A more detailed explanation is offered below.

The CPS is a monthly survey of 60,000 households, which ask about labor market activities during the previous week. Respondents are surveyed for four consecutive months, they are then not interviewed for 8 months, and then they are reinterviewed for another four months. The interviews conducted on the 4th and 8th months contain additional questions on earnings and hours for jobs worked the previous week. I use data from all months from 2010-2013, which corresponds to the ACS years in my sample. I include only records who worked in the private sector, state government, or local government. This discards federal government workers, and the self-employed, neither

of whom are included in the ACS-LEHD matched sample. Finally, I keep only records for which the dependent variable, "Hourly/Non-hourly status" is neither edited nor allocated.

To impute hourly/non-hourly pay, I fit a fully interacted least squares model with a LASSO penalty. The LASSO provides a parsimonious model for both covariate shrinkage and subset selection for an OLS model (Tibshirani, 1996). I use the LASSO in this setting primarily as a tool for subset selection, using both five fold cross validation and the Akaike information criterion (AIC) for selecting the LASSO parameter, which effectively chooses the non-zero covariates in the model. Both methods return similar results, with the best model including an indicator for whether a respondent has a Bachelor's degree or more, and an indicator for whether weekly earnings on the main job is in the top tercile of earnings, as well as their interaction.

For the final imputation model I run a logit model interacting occupation, industry, and tercile of weekly earnings. Although the LASSO indicated that a variable for Bachelor's degree or higher should be included, I omit it from the final imputation model for two reasons. First, although highly correlated with non-hourly pay, there is no *a priori* reason for its inclusion. Unlike weekly earnings, education is not part of the duties test for exemption from overtime. Second, although correlated, I would like to evaluate education separately in the statistical analysis. Including it in the non-hourly imputation would make the results hard to identify and interpret. Finally, I include NAICS industry sectors and major occupation groups in the final model. Job duties is one of the major tests for exemption from overtime, which correlates highly with non-hourly status. Various industries carve out exemptions for overtime and determine pay norms, which argues for its inclusion.

After fitting the model on the CPS, I attach the predicted probabilities for each cell to the final analysis sample. Attaching major occupation groups and industry is straight

forward. I do not observe weekly earning for either the ACS or the LEHD. I calculate annual earnings terciles based on the LEHD, and use that as the tercile to which I map weekly earnings. This assumes the same weekly earnings is earned each week, which scales weekly earnings to an annual earnings measure.[1] I then bin each observation in the final analysis sample by quartile of their likelihood of non-hourly pay.

Summary statistics for the analysis sample by quartile of non-hourly pay are available in Table A.3. Based on prior knowledge and casual observation, the results are largely what one would expect. Observations in the highest quartile of probability they are not paid by the hour have much higher reported hours worked compared to hours paid averaging 14.3% assuming 52 weeks worked. In contrast, quartiles one through three are relatively uniform with hours worked exceeding hours paid by 6.8%, 4.0%, and 6.3% for quartiles one through three, respectively. The remaining stratifying variables change by quartile as expected. Workers in the top quartile are much more likely to be white, male, and have a Bachelor's degree.

A.4 Additional Tables and Figures

[1] Recall the final analysis sample is only for full-year workers.

Table A.1: The Effect of the Unemployment Rate on Hours Worked and Hours Paid

	(1)	(2)	(3)	(4)
	Panel A. Dependent Variable: Log Annual ACS Hours			
Unemployment rate (β)	-0.00126	-0.000917	-0.00206**	-0.00213**
	(0.00110)	(0.00129)	(0.000820)	(0.000851)
	Panel B. Dependent Variable: Log Annual ACS Hours, Winsorized 5%			
Unemployment rate (β)	-0.000763	-0.000275	-0.00138*	-0.00146*
	(0.001000)	(0.00116)	(0.000742)	(0.000767)
	Panel C. Dependent Variable: Log Annual LEHD Hours			
Unemployment rate (β)	0.000121	0.000829	-0.000451	-0.000414
	(0.00123)	(0.00129)	(0.000766)	(0.000747)
	Panel D. Dependent Variable: Log Annual LEHD Hours, Winsorized 5%			
Unemployment rate (β)	0.000387	0.00123	1.10e-05	2.92e-05
	(0.00109)	(0.00113)	(0.000610)	(0.000604)
State FE	X	X	X	X
Commuting Zone Time Trends		X	X	X
Firm & Job controls			X	X
Demographic controls				X

Notes: $N = 218,000$, with 58 commuting zones. Dependent variable indicated by panel title. All regressions estimated using ordinary least squares using the specification outlined in equation 1.2. Cluster-robust standard errors clustered by commuting zone. Stars on standard errors accord to p-values as follows: *** p<0.01, ** p<0.05, * p<0.1.

Table A.2: Firm Size by Likely Exempt Status

	(1)	(2)	(3)
	Non-Supervisory Occ.	*Bottom Half Quartiles*	*Bottom Three Quartiles*
0-19	0.0294***	0.0489***	0.0412***
	(0.0069)	(0.0068)	(0.0054)
20-49	0.00846	0.0190***	0.0138**
	(0.0114)	(0.0072)	(0.0067)
50-249	0.00647	0.00521	0.0009
	(0.0059)	(0.0062)	(0.0050)
250-999	0.0000	0.00371	-0.0046
	(0.0064)	(0.0065)	(0.0053)
1,000-2,499	0.0126	0.00596	0.00232
	(0.0081)	(0.0079)	(0.0069)
	Supervisory Occupations	*Top Half Quartiles*	*Top Quartile*
0-19	0.0599***	0.0615***	0.0591***
	(0.0065)	(0.0064)	(0.00659)
20-49	0.0280***	0.0460***	0.0534***
	(0.0099)	(0.0133)	(0.0163)
50-249	0.0384***	0.0620***	0.0870***
	(0.0058)	(0.0062)	(0.00611)
250-999	0.0198***	0.0457***	0.0690***
	(0.0060)	(0.0068)	(0.00756)
1,000-2,499	0.0202***	0.0641***	0.0902***
	(0.0073)	(0.0089)	(0.0103)
+2,500	0.0391***	0.0607***	0.0823***
	(0.0049)	(0.0060)	(0.00710)
Firm controls	X	X	X
Year-Quarter FE	X	X	X
State FE	X	X	X
Demographic controls	X	X	X
R^2	0.098	0.119	0.124

Notes: $N = 218,000$. Dependent variable in all columns is the difference between log annual ACS hours calculated at 52 weeks and log annual LEHD hours. Each column is its own regression specification. All coefficients reported in reference to largest firm size group (+2,500) interacted with top panel in each of the three regressions. For column (1) that is non-supervisory occupations, column (2) is the bottom two quartiles of probability non-hourly pay, and column (3) is the bottom three quarters of probability non-hourly pay. Cluster-robust standard errors clustered by state employer of dominant job. Stars on standard errors accord to p-values as follows: *** p<0.01, ** p<0.05, * p<0.1

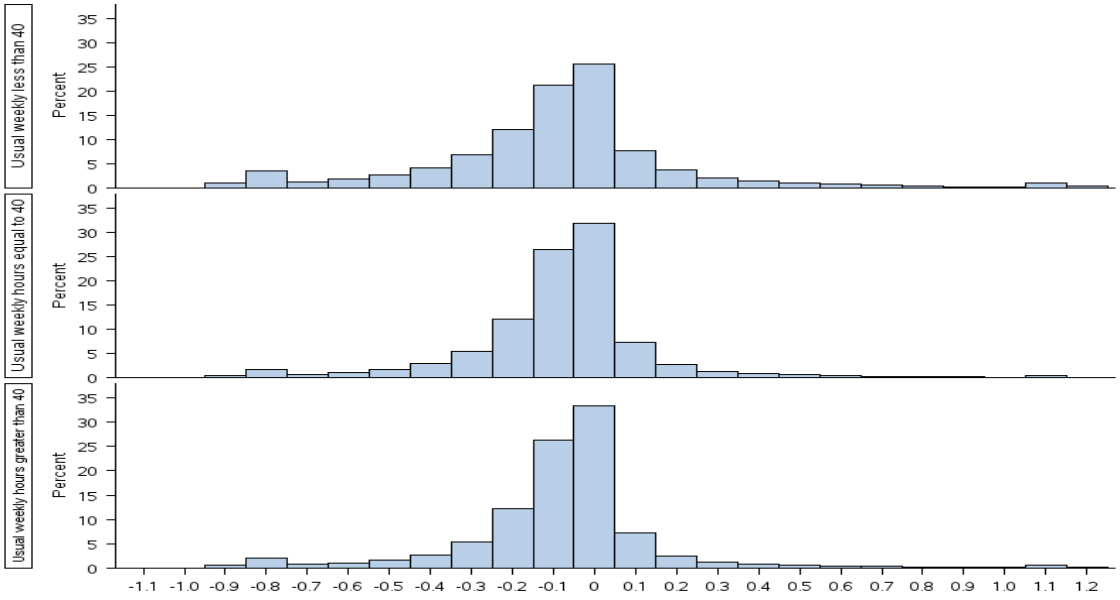


Figure A.1: Distribution of the Difference of Log Labor Earnings (ACS) and Log Labor Earnings (LEHD) by Usual Weekly Hours

Notes: Variable is the difference in log ACS earnings from log LEHD earnings for the full sample, winsorized at the 5% and 95% level. The variable is partitioned by whether an ACS respondent answers that she usually works either 1) less than 40 hours per work (top panel) or 2) exactly 40 hours a week (middle panel) or 3) more than 40 hours per week (bottom panel). Top panel $N = 49,000$, middle panel $N = 108,000$, bottom panel $N = 61,000$. Mean of bottom panel -0.094 [0.318], middle panel -0.077 [0.247] and bottom panel -0.078 [0.248].

Table A.3: Summary Statistics by Quartile of Probability of Non-hourly Pay

Quartile Liklihood Not Paid by Hour	(1)		(2)		(3)		(4)	
	mean	sd	mean	sd	mean	sd	mean	sd
ACS annual hours (52 Weeks)	1,755	575.9	2,040	446.8	2,116	457.9	2,303	401.6
LEHD annual hours	1,690	627.7	1,993	502.5	2,014	481.0	2,006	361.6
Annual hours error (50 weeks, ACS)	0.011	0.301	-0.010	0.238	0.014	0.237	0.099	0.225
Annual hours error (51 weeks, ACS)	0.041	0.284	0.016	0.223	0.039	0.230	0.121	0.222
Annual hours error (52 weeks, ACS)	0.068	0.278	0.040	0.218	0.063	0.228	0.143	0.222
Firm/Job Characteristics								
Unemployment rate (CZ)	7.277	1.956	7.386	1.935	7.377	1.924	7.332	1.802
Private, for-profit firm	0.880	0.325	0.827	0.378	0.586	0.493	0.702	0.457
Likely Exempt occupation (Management)	0.092	0.289	0.199	0.399	0.203	0.402	0.541	0.498
Dominant Job tenure (quarters)	20.07	17.35	25.67	20.12	29.25	21.78	30.89	22.26
Demographic Characteristics								
Age	39.08	15.12	41.69	13.05	43.37	12.11	44.13	11.09
Male	0.455	0.498	0.564	0.496	0.450	0.497	0.575	0.494
Non-white	0.307	0.461	0.248	0.432	0.222	0.416	0.197	0.398
Hispanic	0.090	0.286	0.079	0.269	0.053	0.224	0.033	0.179
Bachelors degree or higher	0.068	0.252	0.122	0.327	0.365	0.482	0.674	0.469
Observations	33,000		60,000		57,000		67,000	

Notes: $N = 218,000$. Annual hours error is the difference between log hours worked in the ACS and log hours paid from the LEHD. The ACS hours paid measure is defined by multiplying the usual weekly hours by the number of weeks paid in each row. The hourly/non-hourly imputation and the description of the construction of the quartiles provided in appendix A.3.

Table A.4: Covariates in Inverse Probability Weighting Probit

Variable	Description
Gender	Indicator for whether male.
Age	16-19, 20-34, 35-49, 50-65, 65+
Education	Less than high school, High School, Some College, BA+
White	Indicator for whether race is white
Hispanic	Indicator for hispanic ethnicity
Citizen	Indicator for whether U.S. citizen
Married	Indicator for whether married
Kids	Indicator for presence of own children
Moved	Indicator for whether moved in last year
Disability	Indicator for whether has a disability
English	Indicator for whether speaks English ”Very well” or ”Well”
Labor Force	Indicator for whether in labor force

Notes: Variables used for reweighting sample weights to account for PIKs missing at random. All variables are indicator variables unless otherwise noted. Construction and use of the weights is described in appendix A.2.

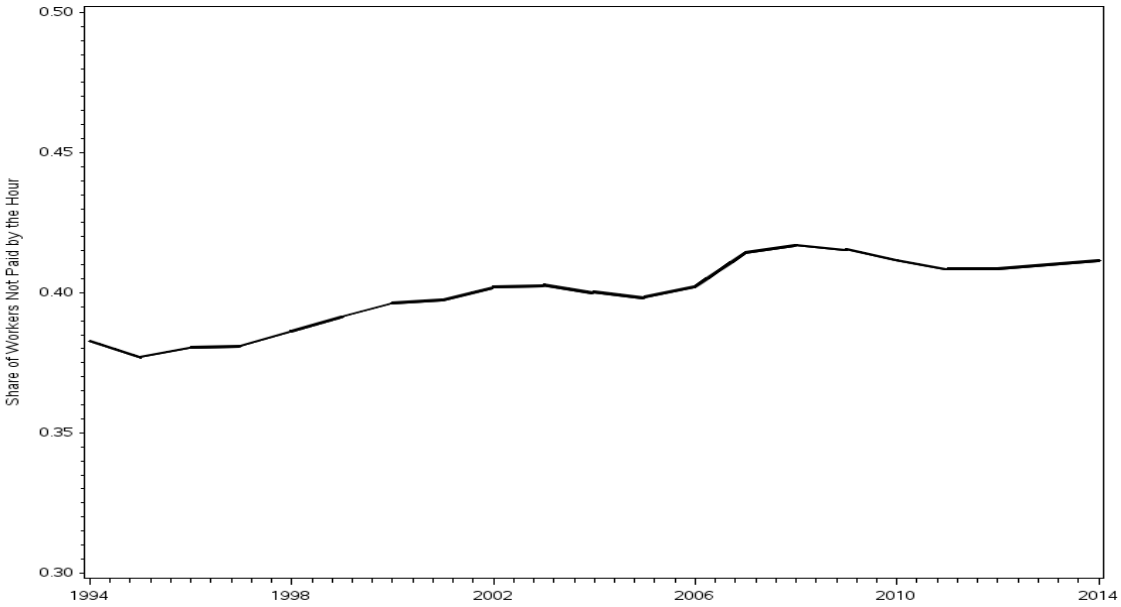


Figure A.2: Share of Wage and Salary Workers not Paid by the Hour, 1994-2015

Notes: Author's Analysis of Current Population Survey, Outgoing Rotation Group data. Sample excludes self-employed and those who worked without pay.

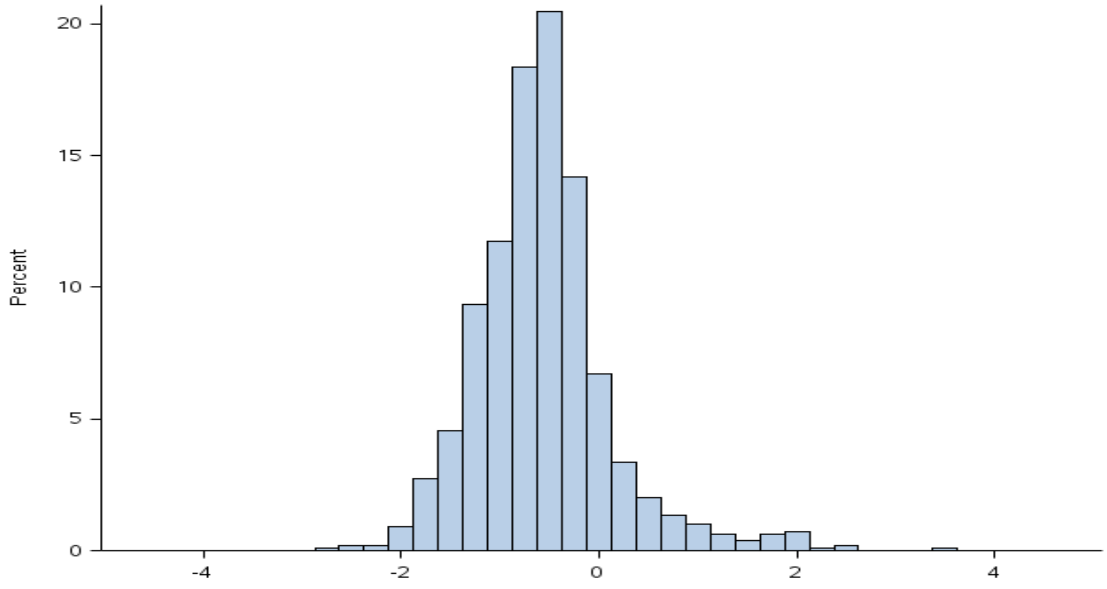


Figure A.3: Distribution of Year-over-year Change in Quarterly Unemployment rates by Commuting Zone

Notes: N=218,000. Mean −0.568 [0.689].

Table A.5: Regression Results for Unemployment Rate and Work Off the Clock, Heterogeneous Effects Subsets

	(1) Supervisory	(2) Non-Supervisory	(3) Top half likely paid non-hourly	(4) Bottom half likely paid non-hourly	(5) At least B.A. degree	(6) Less than B.A. degree
Unemployment rate (β)	-0.00097	-0.00213*	-0.00107	-0.00287*	0.00000	-0.00231**
	(0.00138)	(0.00112)	(0.00120)	(0.00151)	(0.00126)	(0.00099)
Observations	74,000	145,000	131,000	87,000	75,000	143,000
R^2	0.133	0.070	0.129	0.096	0.125	0.075

Notes: $N = 218,000$, with 58 commuting zones. Dependent variable is the difference between log annual ACS hours calculated at 52 weeks and log annual LEHD hours. Supervisory, Column (1), refers to the compliment of non-supervisory. The definition of non-supervisory or production worker comes from the Bureau of Labor Statistics. Column (2), Non-supervisory, is an indicator for observations who meet the definition of a production or non-supervisory worker according to the occupation and industry of her dominant job. "Bottom half likely paid non-hourly", Column (3), refers to observations who are below the median in likelihood they are not paid by the hour. "Top half likely paid non-hourly", Column (4), refers to observations above the median likelihood not paid by the hour. B.A refers to Bachelor's degree. All regressions run using OLS with the same specification as Table 1.2 column (4), subset to cell listed at the top of each column. Cluster-robust standard errors clustered by commuting zone. Stars on standard errors accord to p-values as follows: *** p<0.01, ** p<0.05, * p<0.1.

Table A.6: Regression Results for Firm Growth and Work Off the Clock

					Single Establishment Firms	
	(1)	(2)	(3)	(4)	(5)	(6)
Year/Year Firm Growth	0.00511				0.00585	
	(0.00465)				(0.00531)	
Year/Year Firm Growth Bins						
Low		0.00523				0.00127
		(0.00359)				(0.00314)
Mid-low		-0.00170				-0.00593*
		(0.00327)				(0.00310)
Mid-high		-0.00612*				-0.00930***
		(0.00354)				(0.00298)
High		0.00362				-0.000451
		(0.00374)				(0.00317)
Year/Year Establishment Growth			-0.00510			
			(0.00331)			
Year/Year Establishment Growth Bins						
Low				0.00369		
				(0.00342)		
Mid-low				-0.00461		
				(0.00337)		
Mid-high				-0.00565*		
				(0.00317)		
High				-0.000300		
				(0.00353)		
Observations	218,000	218,000	218,000	218,000	126,000	126,000
R^2	0.102	0.102	0.102	0.102	0.114	0.115

Notes: $N = 218,000$. Dependent variable is the difference between log annual ACS hours calculated at 52 weeks and log annual LEHD hours. Cluster-robust standard errors clustered by state-firm. All year/year growth measures calculated using Haltiwanger et al. (1996). Firm growth calculated using state-firm employment counts in ACS interview quarter and one year prior. Establishment growth rates use the modal establishment from the LEHD unit-to-worker imputation. Firm and establishment growth bins calculated as evenly spaced quintiles of the analysis sample according to firm and establishment growth rates, respectively. Firm/Establishment growth rate bins interpreted in relation to middle quintile, which has mean approximately zero. Columns (5) and (6) subset the sample to only single establishment firms negating use of unit-to-worker imputation. All results estimated with OLS and include firm, job, demographic, and local labor market controls. Stars on standard errors accord to p-values as follows: *** p<0.01, ** p<0.05, * p<0.1.

APPENDIX B

APPENDIX: HOURS ADJUSTMENTS

B.1 Transitions

B.1.1 Full-Time Employment

Hires from Employment to Full-Time Employment

$$\text{hire}^{s,\text{in}}_{i,j,t}(\text{FT}|\text{E}) = \begin{cases} & y_{i,j,t-2} = 0 \text{ and } y_{i,j,t-1} > 0 \text{ and } y_{i,j,t} > 0 \text{ and } y_{i,j,t+1} > 0 \\ 1 \quad \text{if} & y_{i,j',t-2} > 0, \text{ for some } j' \neq j \\ & h_{i,j,t} \geq 400 \\ 0 \quad \text{otherwise} & \end{cases}$$

Hires from Nonemployment to Full-Time Employment

$$\text{hire}^{s,\text{in}}_{i,j,t}(\text{FT}|\text{NE}) = \begin{cases} & y_{i,j,t-2} = 0 \text{ and } y_{i,j,t-1} > 0 \text{ and } y_{i,j,t} > 0 \text{ and } y_{i,j,t+1} > 0 \\ 1 \quad \text{if} & y_{i,j',t-2} = 0, \forall j' \\ & h_{i,j,t} \geq 400 \\ 0 \quad \text{otherwise} & \end{cases}$$

Stayer who transitions from Full-Time Employment to Full-Time Employment

$$\text{stayer}^{s,\text{in}}_{i,j,t}(\text{FT}|\text{FT}) = \begin{cases} & y_{i,j,t-2} > 0 \text{ and } y_{i,j,t-1} > 0 \text{ and } y_{i,j,t} > 0 \text{ and } y_{i,j,t+1} > 0 \\ 1 \quad \text{if} & h_{i,j,t-1} \geq 400 \\ & h_{i,j,t} \geq 400 \\ 0 \quad \text{otherwise} & \end{cases}$$

Stayer who transitions from Part-Time Employment to Full-Time Employment

$$\text{stayer}^{s,\text{in}}_{i,j,t}(\text{FT}|\text{PT}) = \begin{cases} & y_{i,j,t-2} > 0 \text{ and } y_{i,j,t-1} > 0 \text{ and } y_{i,j,t} > 0 \text{ and } y_{i,j,t+1} > 0 \\ 1 \quad \text{if} & h_{i,j,t-1} < 400 \\ & h_{i,j,t} \geq 400 \\ 0 \quad \text{otherwise} \end{cases}$$

B.1.2 Part-Time Employment

Hires from Employment to Part-Time Employment

$$\text{hire}^{s,\text{in}}_{i,j,t}(\text{PT}|\text{E}) = \begin{cases} & y_{i,j,t-2} = 0 \text{ and } y_{i,j,t-1} > 0 \text{ and } y_{i,j,t} > 0 \text{ and } y_{i,j,t+1} > 0 \\ 1 \quad \text{if} & y_{i,j',t-2} > 0, \text{ for some } j' \neq j \\ & h_{i,j,t} < 400 \\ 0 \quad \text{otherwise} \end{cases}$$

Hires from Nonemployment to Part-Time Employment

$$\text{hire}^{s,\text{in}}_{i,j,t}(\text{PT}|\text{NE}) = \begin{cases} & y_{i,j,t-2} = 0 \text{ and } y_{i,j,t-1} > 0 \text{ and } y_{i,j,t} > 0 \text{ and } y_{i,j,t+1} > 0 \\ 1 \quad \text{if} & y_{i,j',t-2} = 0, \forall j' \\ & h_{i,j,t} < 400 \\ 0 \quad \text{otherwise} \end{cases}$$

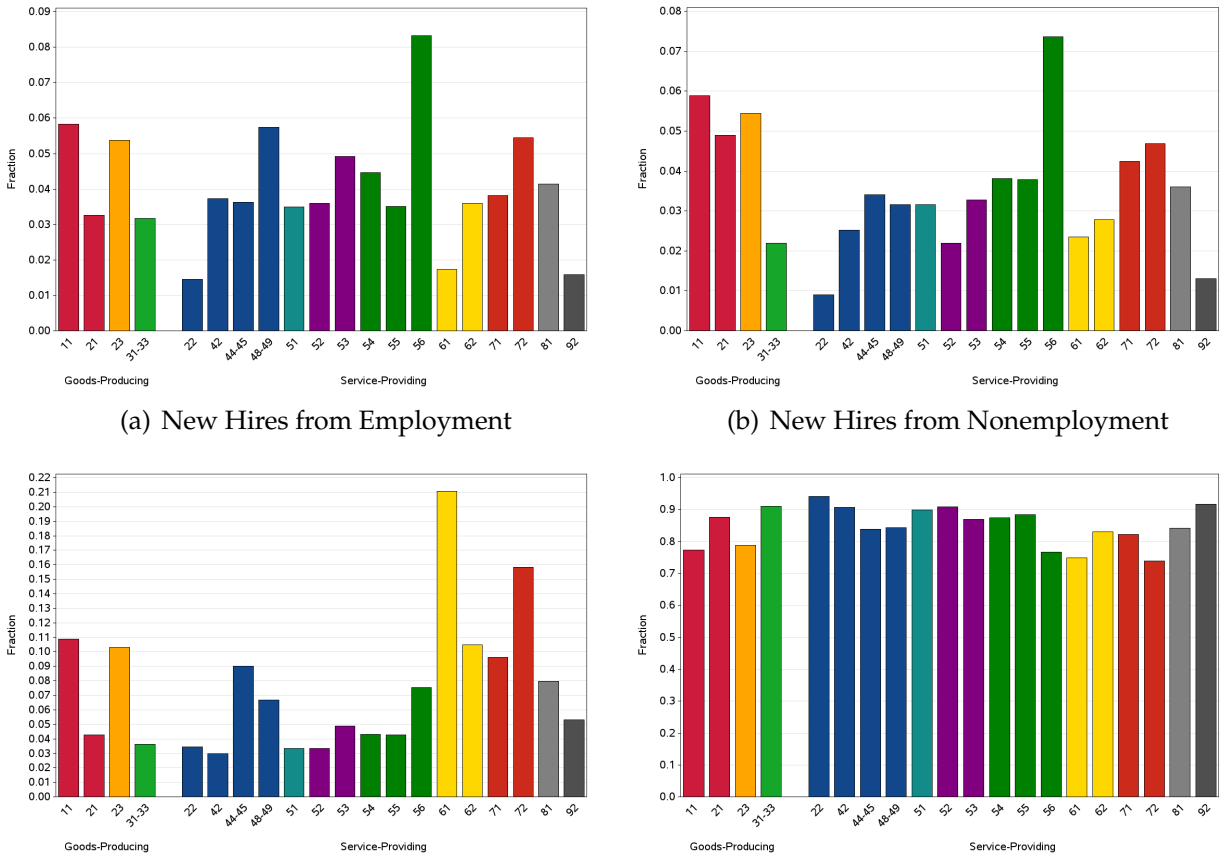
Stayer who transitions from Full-Time Employment to Part-Time Employment

$$\text{stayer}^{s,\text{in}}_{i,j,t}(\text{FT}|\text{FT}) = \begin{cases} & y_{i,j,t-2} > 0 \text{ and } y_{i,j,t-1} > 0 \text{ and } y_{i,j,t} > 0 \text{ and } y_{i,j,t+1} > 0 \\ 1 \quad \text{if} & h_{i,j,t-1} \geq 400 \\ & h_{i,j,t} < 400 \\ 0 \quad \text{otherwise} \end{cases}$$

Stayer who transitions from Part-Time Employment to Part-Time Employment

$$\text{stayer}^{s,\text{in}}_{i,j,t}(\text{FT}|\text{PT}) = \begin{cases} & y_{i,j,t-2} > 0 \text{ and } y_{i,j,t-1} > 0 \text{ and } y_{i,j,t} > 0 \text{ and } y_{i,j,t+1} > 0 \\ 1 \quad \text{if} & h_{i,j,t-1} < 400 \\ & h_{i,j,t} < 400 \\ 0 \quad \text{otherwise} & \end{cases}$$

B.2 Appendix Figures



(a) New Hires from Employment

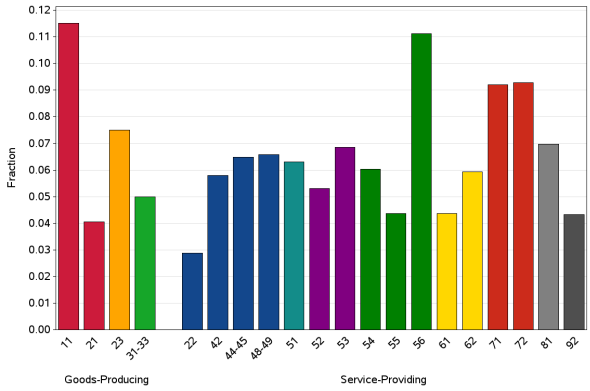
(b) New Hires from Nonemployment

(c) Stayers: Transitions from PT to FT

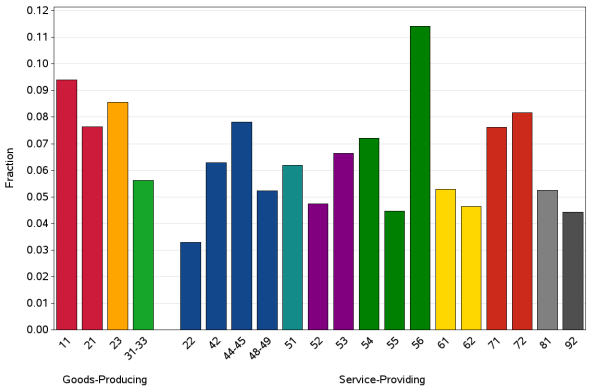
(d) Stayers: Transitions from FT to FT

Figure B.1: Transitions to Full-Time Employment by Industry

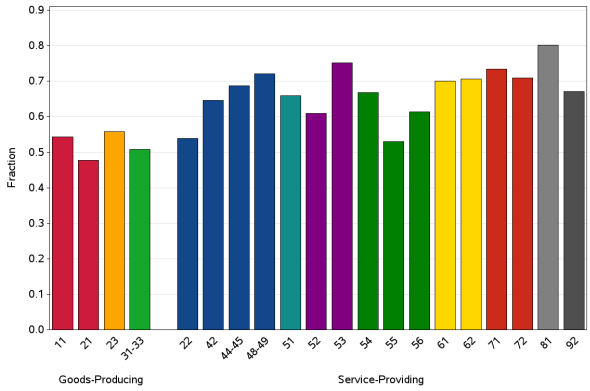
Notes: Subplot (a) plots the share of full-time employment that are hires from employment. Subplot (b) plots the share of full-time employment that are hires from nonemployment. Subplot (c) plots the share of full-time employment that are transitions from part-time employment. Subplot (d) plots the share of full-time employment that are transitions from full-time employment.



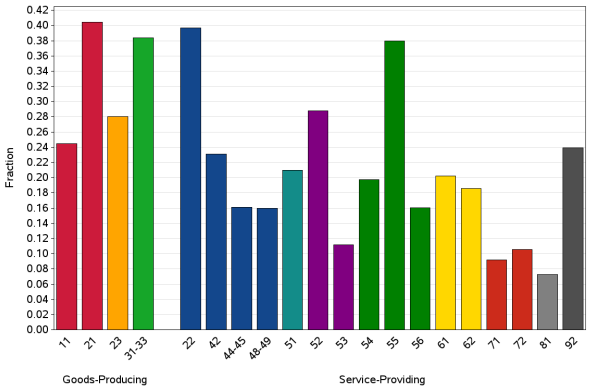
(a) New Hires from Employment



(b) New Hires from Nonemployment



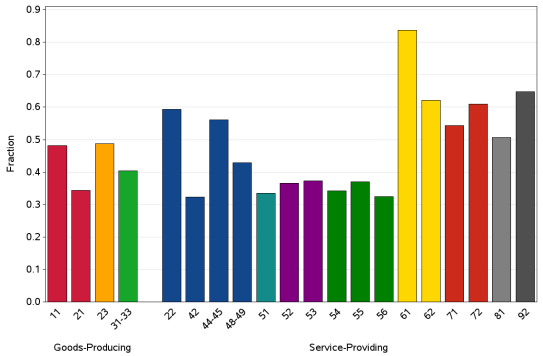
(c) Stayers: Transitions from PT to PT



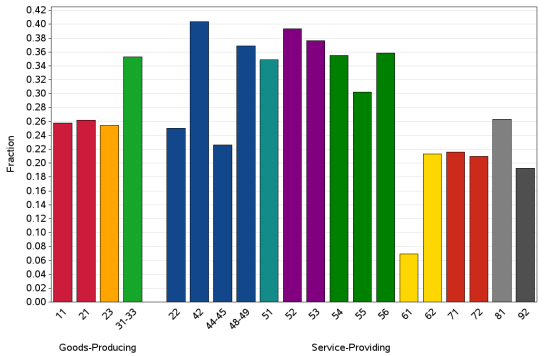
(d) Stayers: Transitions from FT to PT

Figure B.2: Transitions into Part-Time Employment by Industry

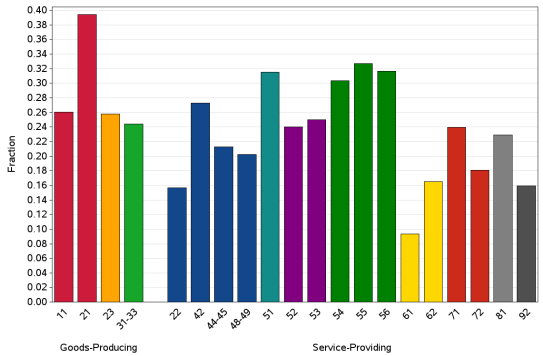
Notes: Subplot (a) plots the share of part-time employment that are hires from employment. Subplot (b) plots the share of part-time employment that are hires from nonemployment. Subplot (c) plots the share of part-time employment that are transitions from part-time employment. Subplot (d) plots the share of part-time employment that are transitions from full-time employment.



(a) PT to FT Transitions

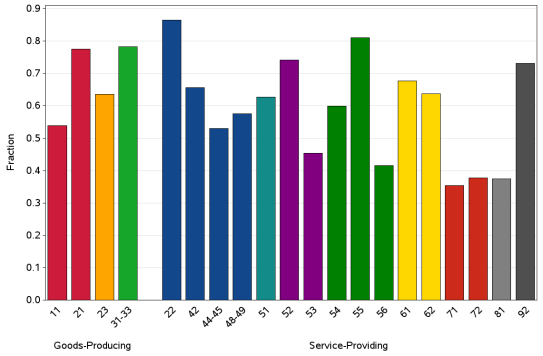


(b) New FT Hires from Employment

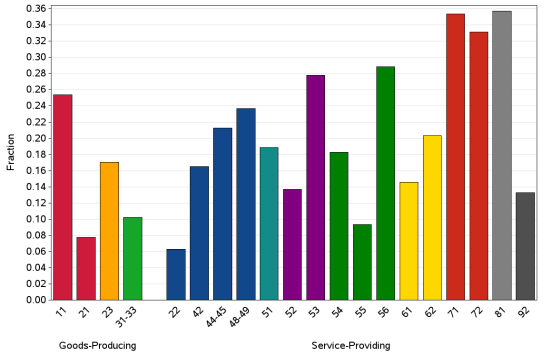


(c) New FT Hires from Nonemployment

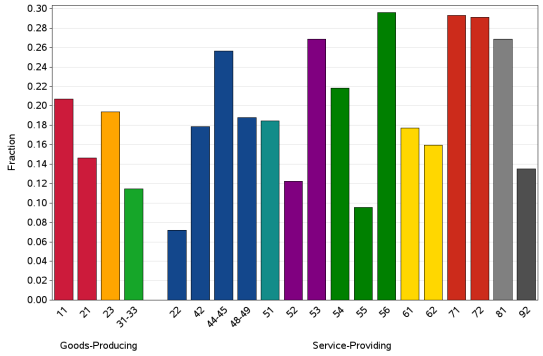
Transitions to New Full-Time Jobs



(d) FT to PT Transitions



(e) New PT Hires from Employment



(f) New PT Hires from Nonemployment

Transitions to New Part-Time Jobs

Figure B.3: Transitions into New Full-Time and Part-Time Jobs

Notes: The top row decomposes new full-time jobs by their origin employment state. Subplot (a) plots the share of new full-time jobs that are within job hours changes from part-time to full-time. Subplot (b) plots the share of new full-time jobs that are new hires from employment. Subplot (c) plots the share of new full-time jobs that are new hires from nonemployment. The bottom row decomposes new part-time jobs by their origin employment state. Subplot (d) plots the share of new part-time jobs that are within job hours changes from full-time to part-time. Subplot (e) plots the share of new part-time jobs that are new hires from employment. Subplot (e) plots the share of new part-time jobs that are new hires from nonemployment.

B.3 Appendix Tables

Table B.1: BLS NAICS Supersectors

GOODS-PRODUCING	
Natural resources and mining	
11	Agriculture, Forestry, Fishing and Hunting
21	Mining
Construction	
23	Construction
Manufacturing	
31-33	Manufacturing
SERVICE-PRODIVING	
Trade, transportation, and utilities	
22	Utilities
42	Wholesale Trade
44-45	Retail Trade
48-49	Transportation and Warehousing
Information	
51	Information
Financial activities	
52	Finance and Insurance
53	Real Estate Rental and Leasing
Professional and business services	
54	Professional, Scientific, and Technical Services
55	Management of Companies and Enterprises
56	Administrative Support and Waste Management and Remediation Services
Education and health services	
61	Educational Services
62	Health Care and Social Assistance
Leisure and hospitality	
71	Arts, Entertainment, and Recreation
72	Accommodation and Food Services
Other services	
81	Other Services (except Public Administration)
Government	
92	Public Administration

Source: See `http://www.bls.gov/sae/saesuper.htm`.

Table B.2: NAICS Sectors

Code	Industry Title	Establishments		Workers		Payroll	
11	Agriculture, Forestry, Fishing and Hunting	N/A		N/A		N/A	
21	Mining	28,643	0.40%	903,641	0.81%	61,331,381	1.18%
22	Utilities	17,595	0.25%	651,234	0.58%	58,922,951	1.14%
23	Construction	598,065	8.41%	5,669,623	5.05%	272,546,238	5.26%
31-33	Manufacturing	297,191	4.18%	11,214,165	9.99%	593,397,004	11.44%
42	Wholesale Trade	419,464	5.90%	5,881,913	5.24%	362,121,460	6.98%
44-45	Retail Trade	1,062,083	14.94%	14,703,529	13.10%	369,001,350	7.12%
48-49	Transportation and Warehousing	213,809	3.01%	4,305,464	3.84%	183,840,920	3.55%
51	Information	138,341	1.95%	3,321,226	2.96%	269,069,624	5.19%
52	Finance and Insurance	468,183	6.58%	6,040,880	5.38%	523,553,280	10.10%
53	Real Estate Rental and Leasing	354,106	4.98%	1,923,770	1.71%	85,325,956	1.65%
54	Professional, Scientific, and Technical Services	856,463	12.05%	8,203,735	7.31%	581,406,127	11.21%
55	Management of Companies and Enterprises	53,765	0.76%	3,082,292	2.75%	308,764,261	5.95%
56	Administrative Support and Waste Management and Remediation Services	386,387	5.43%	9,774,262	8.71%	329,013,077	6.35%
61	Educational Services	67,960	0.96%	642,334	0.57%	17,283,363	0.33%
62	Health Care and Social Assistance	831,303	11.69%	18,414,757	16.40%	801,239,522	15.45%
71	Arts, Entertainment, and Recreation	124,591	1.75%	2,081,668	1.85%	64,051,613	1.24%
72	Accommodation and Food Services	662,489	9.32%	12,007,689	10.70%	196,103,341	3.78%
81	Other Services (except Public Administration)	529,691	7.45%	3,430,711	3.06%	108,185,736	2.09%
92	Public Administration	N/A		N/A		N/A	

Notes: Establishment and employee counts from the 2012 Economic Census (ECN_2012_US_00A1). The establishment count is the number of employer establishments. The worker count is the number of paid employees for the pay period including March 12. Payroll is annual payroll in $1,000.

APPENDIX C

APPENDIX: TOTAL VARIABILITY

C.1 Details of the Methodology for Imputing Missing Birth Date, Sex, Race, Ethnicity, and Education[1]

The LEHD data come from state UI systems' reports of a worker, a firm, and the worker's quarterly earnings. The data the Census Bureau receives from the states contain no information on worker characteristics including age, sex, race, ethnicity, and education. These individual characteristics are a unique attribute of the QWI and LODES. In order to provide the individual characteristics, the Census Bureau attaches its own surveys as well as administrative data from other U.S. government agencies to the LEHD UI data. In cases where the outside surveys and administrative data are not sufficient to account for all characteristics for all workers, the characteristics are imputed.

This appendix documents the methodology for imputing missing individual characteristics in the LEHD infrastructure files. The appendix describes the outside data sources that provide the individual characteristics that form the basis of the imputation. The candidate imputation models and the basis for their selection are also documented. After explaining the monotone missing data pattern and the final implementation of the imputation process, the quality of the imputation is assessed. At the end of the process, the complete set of individual characteristics is stored in the Individual Characteristics File (ICF), which stores the individual characteristics for all workers who appear in the LEHD UI data including 10 draws of the imputation model for each characteristic that is imputed.

[1]Portions of this appendix are based on an unpublished technical memo dated February 1, 2011 by John Abowd, Henry Hyatt, Mark Kutzbach, Erika McEntarfer, Kevin McKinney, Michael Strain, Lars Vilhuber, and Chen Zhao.

The main source data for race and ethnicity is the 2000 Decennial Census of Population and Housing (short form). For birth date and sex, the Census Numident – Census Bureau version of the Social Security Administration's (SSA) Social Security Number (SSN) master database – is the only source. In cases where the race and ethnicity data are incomplete (i.e. an individual's response to the 2000 Census or ACS was not available) an imputation of an individual's race and ethnicity category was computed conditional on the limited race and ethnicity information available in the Census Numident file (if available). The source data for education is the 2000 Decennial Census of Population and Housing Sample Data (long form). Since education is dynamic, particularly for young workers, education data are only imputed for workers aged 25 and older.

The missing characteristics are imputed using a Bayesian version of the continuous-discrete multivariate product kernel density (KDE) approach. In some instances a multinomial model with Dirichlet priors was employed. These missing data follow a monotone pattern. The characteristics are imputed in three stages, with data completed from the previous stage used in the imputation model for the next stage. The end results is 10 implicates of completed data drawn from estimates of the posterior distribution of the characteristics.

To assess the out-of-sample performance of the imputation model, two separate tests are used. First, the completed race, ethnicity, and education variables were matched to a sample of respondents from the ACS (2000-2010). These comparisons show highly accurate imputation rates, particularly for the larger race and ethnicity groups: White (95% accuracy), Black (90% accuracy), Asian (85% accuracy), and Hispanic (80% accuracy). For education, the results are adequate, but they do not display the same level of accuracy.

In addition to conducting ACS comparisons, the geographic variability captured by our education model was also assessed. Using a sub-sample of workers who have a recorded 2000 Decennial Census (long form) education response, tabulations

of beginning-of-quarter employment, full-quarter employment, and average quarterly wages for full-quarter employees by both the actual and imputed value are calculated. These comparisons show close correspondence, particularly for wages. At the statewide level, the difference between full-quarter wages within education categories for reported and imputed education ranges from -6.8% to +8.0% with some cells within 0.2%. The share of beginning of quarter employment in each education category varies by a range of -5.3 to 6.6 percentage points with most cells within 2 percentage points.

The rest of this appendix proceeds as follows. Section C.1.1 describes the selection of the missing data model for imputing the individual characteristics, Section C.1.2 details the implementation of the models for each of the characteristics, and Section C.1.3 assesses the quality of the imputation.

C.1.1 Methodological Approach

Missing, birth date, sex, race, ethnicity, and education were imputed using multiple imputation following Rubin (1987). The candidate imputation models were implemented and tested before selecting a final procedure at each stage of the imputation. We compared several different estimators: (i) the standard Li and Racine (2003) mixed continuous-discrete KDE (LR); (ii) a Bayesian Li-Racine method based on an approach developed by Zhang et al. (2006) for estimating the posterior of the bandwidth parameter (ZH); (iii) a multinomial distribution with a Dirichlet prior combined with Bayesian bootstrap re-sampling (BB); (iv) a cold deck (the equivalent of hot deck methods when all the data are given) (CD); and (v) a naïve method (modal imputations in sub groups) (NA).

To assess the performance of each candidate, a 3-dimensional distribution for birth year, race/ethnicity, and education was created using data from the Current Population

Survey (CPS).[2] Using balanced half-sample cross validation, the research question examined was: with 100% imputation rates, what are the Kullback-Leibler Divergence (KL) and Mean Squared Error (MSE) losses associated with each of these methods, assuming ignorable missing data.

The combined CPS's were treated as a synthetic population of 170,000 individuals. For each of the candidates the KL and MSE criteria were estimated using the CPS data. The KL was computed by comparing the actual and imputed distributions. Half-samples were created randomly by assigning in-scope individuals permanently to A and B subpopulations of equal sizes. All models were fit on sub-population A, then used to impute sub-population B, subsequently the process was reversed with the estimates based on the B sample used to impute A. Hence, every member of the population received imputed values for every model based on an out-of-estimation-sample forecast. All the estimators were compared for a variety of stratifying schemes. KL and MSE performances were considered when adopting strategies for choosing stratifiers used in the final implementation.

The ZH and LR methods underestimated the KL and MSE losses, using BB as the standard, but often by less than 10%. In many cases, the ZH and LR methods were effectively indistinguishable from the BB. LR, ZH and BB substantially out-performed both the cold-deck and naïve models. Up to two levels of stratifiers, with a total of eight subpopulations, were tested.[3] There were large (one or two orders of magnitude) improvements in the KL and MSE loss estimates as stratifiers were added. The BB, ZH, LR, and CD methods all led to the same conclusions about which stratifiers to consider first, and to the conclusion that with subpopulations of 20,000 from a population of 170,000, all stratifiers improved the KL and MSE measurably. The NA model performed poorly, which

[2]Specifically, the 1998 through 2005 pooled March data.

[3]This approximately evenly stratified the CPS population into sub-populations of about 20,000 records each.

was expected. The BB, ZH, and LR models all outperformed the CD, and were roughly comparable.

LR and ZH methods were implemented for birth date, sex, race and ethnicity, and partly for education. A variant of BB was also implemented for education. The two KDE methods perform well relative to BB, directly handle continuous data, and allow greater flexibility in the actual implementation. Occasionally the cells created by the stratifiers became too small to estimate with the KDE methods necessitating the use of BB.

C.1.2 Implementation

The missing data follow a special monotone pattern, allowing us to complete the data in three stages. Birth date, sex and place of birth (completed but not used in any tabulations) have the least missing data (about 5% of cases), and are (almost) always missing if race, ethnicity or education are missing. Race and ethnicity are missing for about 18% of the individuals, and are always missing if education is missing. The variables with the least amount of missing data first (sex, birth date, and place of birth), were imputed first. Missing race and ethnicity were imputed next, taking the imputed values for birth date, sex, and place of birth as given. Finally, missing education was imputed.[4]

At each stage, the variables imputed in the previous stage(s) along with various detailed work history, firm, and co-worker characteristics derived from the unemployment insurance wage data were used to create cells. The design of this stratification scheme was based on the tests described above using the CPS test synthetic population.

The models are fit using persons with complete information at each stage with a full

[4]The monotone missing data pattern is a result of the process by which SSNs are attached to the 2000 Decennial. Sex, date of birth, and place of birth are available on the Census Numident. These data are virtually complete because they are necessary for the administration of the program. Only valid SSNs can be attached to a given 2000 Decennial record, generating the monotone missing data pattern.

set of interacted explanatory variables. Intuitively, the models partition observations by stratifying variables (workers) into cells, and then estimate the distribution of interest for each cell. For example, a model for education would estimate the education distribution for a cell of white women ages 35-44 with non-missing education. Observations who are white women ages 35-44, and who are missing education would then receive 10 draws from the distribution fit on that cell.

Birth date, Sex, and Place of Birth

The Social Security Administrations Numident is the source for birth date and sex. The Numident is the Social Security Administrations master file of issued SSNs, which contains a near universe of birth date and sex information of U.S. workers. Approximately 97% of workers in the LEHD data can be matched to the Numident. Birth date and sex are multiply imputed for approximately 7% of records.

A non-parametric KDE is used to estimate the joint distribution of sex and age conditional on various observed characteristics. The model is state specific, and uses the complete set of yearly earnings and employment indicator variables spanning the entire time a states records are available. The estimated model parameters are used to calculate a predicted probability the record is male. Age is imputed is a similar manner. QWI and LODES report age in eight discrete categories. For the purpose of imputing birth date, a record with missing birth date information is assigned into one of the eight age categories using the KDE model similar to the sex imputation. Date of birth is then assigned based on the distribution of ages within each of the eight age categories for entering workers. As with sex, 10 independent draws assign 10 separate dates of birth for each record contain missing date of birth.

The sex and place of birth variables are unordered categorical, and age is real numeric.

For estimating the distributions, the following stratifiers were used. Stratifiers for stage A:

- Modal place of birth non-native-born coworkers
- Proportion of coworkers that are male (> 50%).
- New worker indicator.

Race and Ethnicity

To implement the race and ethnicity imputation, the following steps were taken. First, since the 2000 Census Short Form provided substantial respondent flexibility for reporting race and ethnicity, it was necessary to simplify the reporting for the imputation models. The vast majority of respondents chose single race and ethnicity categories. A small fraction of the population (less than 3%) reported multiple race and/or ethnicity responses. In compliance with OMB statistical policy, the multiple race responses were collapsed into a single category (two or more races), and ethnicity was collapsed to two responses (Hispanic and not Hispanic). For the respondents who reported "some other race," the actual response was set to missing and they were imputed into one of the OMB-approved race categories.

The non-parametric unordered KDE modeled the joint distribution of race and ethnicity. The model incorporates the imputed age and sex information from the previous step. The race variable is grouped into seven different categories, and the ethnicity variable into just two: Hispanic and not Hispanic. The principal source for race and ethnicity information comes from the 2000 Census decennial short form. Subsequent iterations of the model also incorporate race and ethnicity information from the American Community Survey. Approximately 82% of persons found in the LEHD have valid race and ethnicity information from either the decennial Census of the American Community Survey. For

the remaining records with missing race or ethnicity, the values are multiply imputed.

The ethnicity categories on the QWI tabulations by race and ethnicity are:

1. Hispanic or Latino
2. Not Hispanic or Latino

The race categories on the QWI tabulations by race and ethnicity are:

1. White Alone
2. Black or African American Alone
3. Asian Alone
4. Native Hawaiian or Other Pacific Islander Alone
5. American Indian or Alaska Native Alone
6. Two or More Races.

Race and Ethnicity are both unordered categorical variables. The stratifiers for stage B include both age and place of birth from stage A. In addition, there are:

- Collapsed race/ethnicity cells from the Census Numident
- Average yearly earnings quartiles.
- Coworker fraction white and coworker fraction Hispanic.
- Co-resident fraction white and co-resident fraction Hispanic.

Education

The data for the education imputation come from the 2000 Decennial Census Long Form. Approximately 7% of LEHD workers have valid education information.[5] The modal re-

[5] A recent update includes the ACS after 2000. This increases the number of workers with valid education information to 15%.

sponse "high school graduate, no college" was retained exactly. Three additional categories were created by collapsing the other responses from the 2000 Decennial Census Long Form education variable. The education categories are:

1. Less than a high school diploma
2. High school graduate, no college
3. Some college or Associates degree
4. Bachelor's degree or above.

Unlike race and ethnicity, which were modeled as time-invariant, a person is at risk to accrue additional formal education after entering the workforce, however, this risk declines with age. Individuals generally complete high school before age 20, while Bachelor's degrees are disproportionately attained between the ages of 22 and 25. To ameliorate concerns of younger workers attending post-secondary education, the QWI and LODES only report and impute education data for workers at least age 25.

A Bachelor's degree is almost always required to pursue a graduate degree. Associate degree and some college were collapsed into a single category. The resulting ordered categorical education variable allows the use of an informative kernel when estimating the education density. The stage C stratifiers include the imputed variables from stages A and B as well as:

- Place of birth by income quantile.
- Native and Non-native status.
- Modal NAICS (6 categories) for dominant job.
- Collapsed race and ethnicity cells.
- Coworker fraction male.
- Full-quarter earnings deciles.

- Co-resident fraction white and co-resident fraction Hispanic.

For education, the multinomial-Dirichlet (called BB above, but with no final bootstrap step) was used. Although the LR KDE has improved out of sample performance for imputing education, in the current implementation a fully interacted log-linear model with flat priors was used instead because of its superior performance in small geographic cells. When using stratifiers with a large number of outcomes (detailed geography in particular), the number of cells became too large relative to the sample size. To solve this problem we estimated a log-linear model with a reduced set of parameters. This allows us to include stratifiers as main effects only or with limited interactions, improving overall performance. This is essentially a small-area estimator with mean vector given by the main effects associated with the stratifiers and local effect estimated from the log-linear model.

C.1.3 Quality of the Results

For imputations of race and ethnicity, the chief quality check is a detailed comparison of the completed race and ethnicity variables to a matched sample of respondents on the American Community Survey (ACS). Because the ACS was not used as an input for the imputation models, the ACS provides an out-of-sample performance assessment.

The primary question posed by this analysis was: how frequently does the missing data model impute individuals with no 2000 Census race or ethnicity information to the same race or ethnicity category they indicate in the ACS? The results show very accurate imputations for most race and ethnicity groups, although there is variation across ACS race and ethnicity categories. The highest levels of accuracy, defined here as imputing a response on the LEHD infrastructure consistent with ACS race/ethnicity response, are

for the largest race and ethnicity groups: White (95% accuracy), African-American (90% accuracy), Asian (85% accuracy), and Hispanic (80% accuracy).

Defining an accuracy measure for Native American populations (American Indian, Alaska Native, Native Hawaiian or Pacific Islander) proved more problematic as a matched sample of Census/ACS respondents indicated that a large share of these respondents diverged in their race responses between the Census and the ACS. However, for Native Americans that answer both surveys consistently, imputed LEHD race corresponds to self-reported race well over half of the time. A sizable share of self-reported Native Hawaiians and Pacific Islanders are imputed to Asian in the LEHD infrastructure, in part because a key stratifier for the race imputation (the race variable on the Census Numident) does not separate Pacific Islanders from Asians.

For imputations of education, multiple levels of quality checks were employed. In addition to comparisons with the ACS, a comparison of key QWI variables for three sample states by education, and education x sex, was analyzed using both reported education and imputed education. This analysis uses a sample of workers in the LEHD infrastructure that has a reported 2000 Census long form education response, for which an imputed response was also generated for this assessment. Beginning of quarter employment (B), full-quarter employment (F), and average monthly wages for full-quarter employees (Z_W3) were studied using both respondent-supplied education and imputed education. These indicators were computed for both the reported value of education and for each of the 10 education implicates. The difference between the value of the QWI indicator using reported education and the average value for the indicator using imputed education over the 10 implicates was studied.

For B, F, and Z_W3 analyzing the Education x Sex breakdown at the statewide level, the correspondence is quite close. In statewide Education X Sex tabulations, the difference between average full-quarter wages within categories for reported and imputed educa-

tion ranges from -8.1% to +9.4%. The share of beginning of quarter employment in each education category varies by a range of -5.3 to +6.6 percentage points with the smallest difference being less than 0.001 percentage points at the statewide level. Differences in male/female wage gaps and employment by education across states are largely retained in the imputed results.[6]

ACS Results

To construct the review of imputation quality the results were merged with the ACS. First, three years of person-level data from the ACS were appended together. The same ICF variables used in the imputation were constructed from the unedited responses on the ACS. The education, race, and ethnicity characteristics constructed form the ACS were then merged into the newly created ICF by PIK. Due to the dynamic nature of education, only records older than 25 years of age after April 1st 2000 (according to ICF variable *dob*1) were retained for the analysis.

The ICF records were then stratified for each variable. The records were partitioned by variable according to whether they contained a corresponding valid ACS response. Records were then further subdivided into whether or not the ICF variable was imputed creating four mutually exclusive and collectively exhaustive groups. The group containing records for which there was no corresponding valid ACS variable, and in which the value was not imputed, serves as the baseline distribution for each variable. For the two groups for which a valid ACS response exits – ICF variable imputed, and not imputed – the distribution of the ICF variable was computed conditional on the ACS response for each of the two groups.

In addition to the conditional distribution means, confidence intervals were computed

[6]For disclosure limitation, all results are rounded to three significant digits.

for each value of the distribution using the Rubin methodology (within and between impute variance) to draw confidence intervals around the each category using all implicates of the imputed data. Standard errors are calculated using the following formula (described in U.S. Census Bureau (2003a)),

$$stnd\_error = D\sqrt{\frac{S-1}{B}\left(acc_{pct}\right)\left(1-acc_{pct}\right)} \tag{C.1}$$

where D is the corresponding US design factor for the standard error, S is the number of persons in each of the mutually exclusive categories corresponding to a particular variable minus 1, and B is the population count over age 25 according to the 2000 Decennial Census SF3 file for each category.

For the persons not imputed in the ICF and not matching to the ACS, variable-specific design factors, to account for over-sampling of some populations, were taken from the "Accuracy of Microdata Sample Estimates: Census 2000 PUMS Standard Error Design Factors (U.S. Census Bureau, 2003a)." For Persons matching to the ACS, variable specific design factors were taken from U.S. Census Bureau (2003b).

For each category of each race, ethnicity and education variable, the imputation model was more informative than a random allocation across categories would have been. The models assigned a higher share of individuals to the same category as those persons responded in the ACS than would be expected if the imputation models assigned categories completely at random from the aggregate distribution. The analysis shows, however, that there is considerable variation in imputation quality across variables.

Tables C.1, C.2, and C.3 show the results and 90-10 confidence intervals for the imputation quality analysis for the variables education, ethnicity, and race, respectively. Each table contains the results for each valuable broken out by the individual categories of the variable as reported in the ACS. Table C.1 displays the results for education. The major row heading has the categories for the four possible ACS responses as well as the cate-

gory for ICF records who do not match to a valid education category. The latter group is the first row in the table. The minor row heading for "Not in ACS" indicates that in addition to not matching to the ACS, this group includes only ICF records whose education categories were not imputed. Moving across the first row, the remaining columns give the education distribution for this group. The remaining rows of Table C.1 give the distribution of education conditional on a particular ACS value of education. The minor row headings indicate that these groups are further partitioned by whether the ICF value was imputed.

Table C.1: Distribution of ICF Categories across ACS Response Categories, Education

Distribution of catagories in ICF	90% CI	< High School	High School	Some College	≥ Bachelor
Not in ACS					
Baseline: not imputed	Upper	13.8%	29.6%	30.5%	26.2%
	Mean	13.7%	29.6%	30.5%	26.2%
	Lower	13.7%	29.6%	30.5%	26.2%
ACS: Less than High School					
Impute: imputed, ACS is '< High School'	Upper	26.3%	33.8%	26.9%	14.3%
	Mean	26.0%	33.5%	26.6%	14.0%
	Lower	25.6%	33.1%	26.4%	13.6%
Target: not imputed, ACS is '< High School'	Upper	80.8%	15.0%	4.2%	1.1%
	Mean	80.4%	14.7%	4.0%	1.0%
	Lower	80.0%	14.3%	3.8%	0.9%
ACS: High School					
Impute: imputed, ACS is 'High School'	Upper	14.8%	35.6%	31.6%	18.8%
	Mean	14.6%	35.4%	31.4%	18.6%
	Lower	14.5%	35.1%	31.2%	18.4%
Target: not imputed, ACS is 'High School'	Upper	6.5%	81.5%	12.0%	0.9%
	Mean	6.3%	81.2%	11.7%	0.8%
	Lower	6.1%	80.8%	11.4%	0.7%
ACS: Some College					
Impute: imputed, ACS is 'Some College'	Upper	11.0%	29.9%	33.5%	26.4%
	Mean	10.8%	29.7%	33.3%	26.2%
	Lower	10.7%	29.5%	33.1%	25.9%
Target: not imputed, ACS is 'Some College'	Upper	1.4%	11.3%	85.2%	3.0%
	Mean	1.3%	11.0%	84.8%	2.9%
	Lower	1.2%	10.7%	84.5%	2.7%
ACS: ≥ Bachelors					
Impute: imputed, ACS is '≥ Bachelors'	Upper	6.1%	19.0%	28.6%	47.2%
	Mean	6.0%	18.8%	28.4%	46.9%
	Lower	5.8%	18.6%	28.1%	46.6%
Target: not imputed, ACS is '≥ Bachelors	Upper	0.3%	0.9%	4.8%	94.6%
	Mean	0.2%	0.8%	4.6%	94.3%
	Lower	0.2%	0.7%	4.4%	94.1%

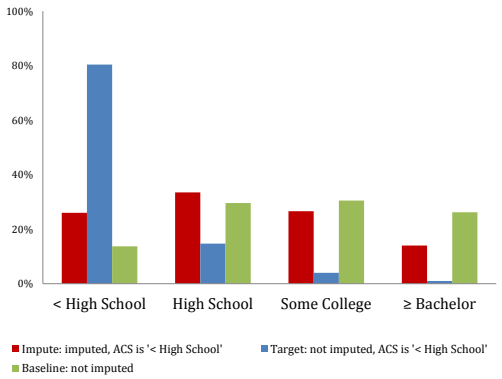
Notes: 90% CI are 90% confidence intervals of the mean. Major row heading is the value of the ACS variable. Minor row heading is the value of the ICF variable. Major row header "Not in ACS" denotes records that did not match to the ACS.

Figure C.1 depicts the two education distributions for each value of education in the

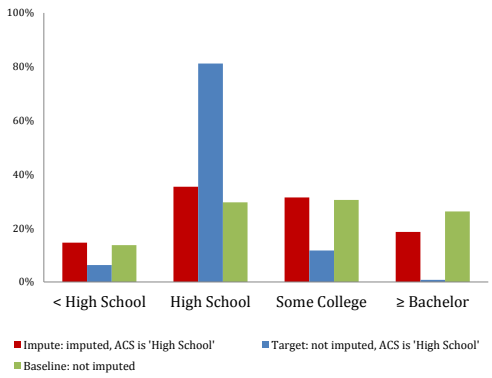
ACS. This depicts graphically what is presented in Table C.1. Each sub-figure corresponds to an ACS value. The blue bars give the distribution of those records, which were not imputed. This serves as the target distribution. The red bar gives the distribution of the records which were imputed. Ideally, this would line up perfectly with the blue bars, but that is not always the case. The green bar shows the overall distribution for records, which were not imputed, and which did not match to the ACS. This is the baseline distribution, and it does not vary across ACS categories. In addition to education, figures depicting impute quality by matching to the ACS are available for race and ethnicity. For each category of each variable, the impute model should not be expected to be much better than the matched ACS responses, so the red line is unlikely to be greater than the green line. The green line does not always equal 1 (or 100%) for the specified ICF category because some people responded differently on the Decennial Census or Numident than they did on the ACS.

The education figures show the most accurate imputations were for the "High School" and "Bachelor's degree and above" categories. The blue line in Figure C.1(d) shows that a little over 94% of records reporting "Bachelor's degree and above" in the 2000 Decennial also reported the same value in the ACS. Of the records imputed into the "Bachelor's degree and above" category and matched to the ACS (red bar), slightly less than 47% had the same value in the ACS. The corresponding values for "High School," Figure C.1(b), are 81.2% (blue bar) and 35.4% (red bar).

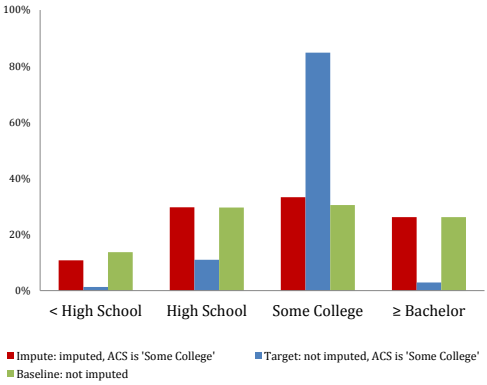
The imputations for the education categories "Less than High School" and "Some College" were somewhat less successful, as measured by correspondence with the ACS. The red bar in Figure C.1(c) gives a rate of 84.8% correspondence between the Decennial and ACS for records which were not imputed and had a value of "Some College." The blue bar depicting correspondence for records which were imputed shows a rate of 33.3%. For "Less than High School" in Figure C.1(a), the two rates are 80.4% (red bar) and 26.0% (blue



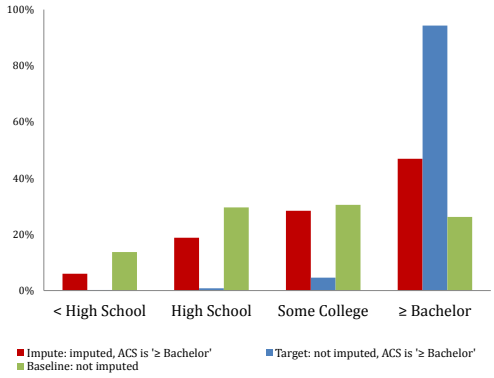
(a) Less than High School



(b) High School



(c) Some College



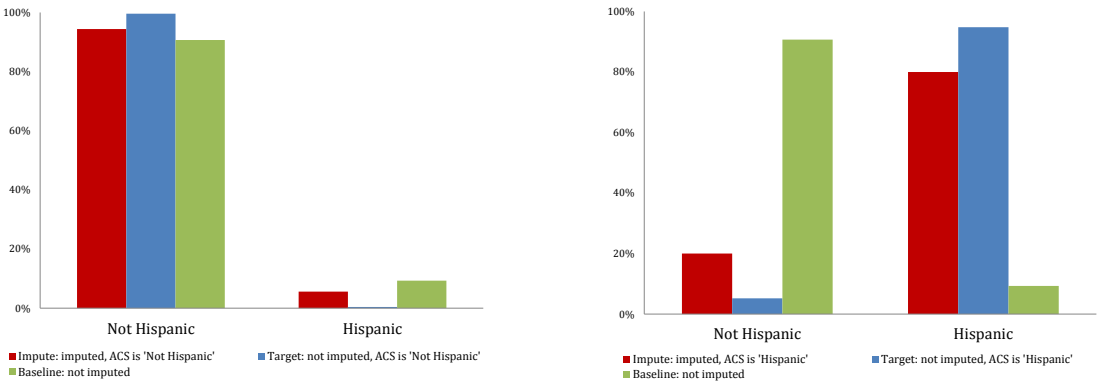
(d) Bachelor's degree and above

Figure C.1: Impute versus Target: Education

Notes: Sub-figure titles correspond to the value of the education variable in the ICF. The blue bars show the distribution of education in the ICF among records that were not imputed. The red bars shows the distribution of education in the ICF among imputed records. The green bars show the distribution of education among records in the ICF that were not imputed *and* did not match to the ACS. The green bars do not vary across sub-figures. See Table C.1 for more detail.

bar). The lower rate of correspondence for all education values compared to "Bachelor's degree and above" are expected, as some Decennial respondents will have completed more schooling upon responding the ACS at a later date.

For ethnicity, the imputation procedure was more accurate than with education. The population for ethnicity is 90.7% "not Hispanic" versus 9.3% "Hispanic" according to the 2000 Decennial. Figure C.2(a) shows that conditional on reporting "not Hispanic" in



(a) Not Hispanic (b) Hispanic

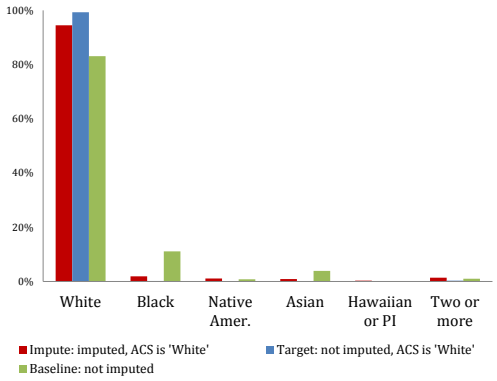
Figure C.2: Impute versus Target: Ethnicity

Notes: Sub-figure titles correspond to the value of the ethnicity variable in the ICF. The blue bars show the distribution of ethnicity in the ICF among records that were not imputed. The red bars shows the distribution of ethnicity in the ICF among imputed records. The green bars show the distribution of ethnicity among records in the ICF that were not imputed *and* did not match to the ACS. The green bars do not vary across sub-figures. See Table C.2 for more detail.

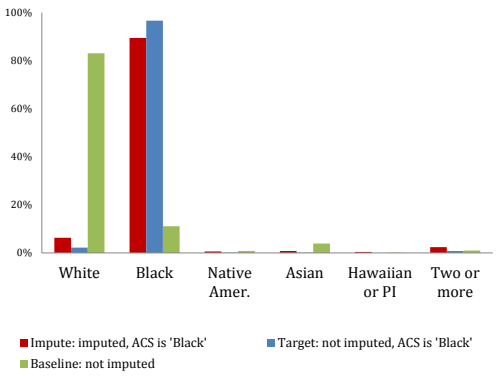
the ACS, approximately 94.4% are imputed into the "not Hispanic" group compared to 99.6% of ACS respondents who were not imputed and report being "not Hispanic" in the Decennial Census as well as the ACS. For the Hispanic group, depicted in Figure C.2(b), these numbers are 80.0% and 94.8%, respectively.

For race, results vary by ACS category. White, Black, and Asian have highly accurate imputations. For these groups, the results are depicted in Figure C.3. For White, Black, and Asian, the rates imputed into those categories conditional on the same ACS response is 94.5%, 89.5%, 83.7%, respectively. This shows relatively high quality as the target distributions are 99.3%, 96.7%, and 94.5%, for White, Black, and Asian, respectively.

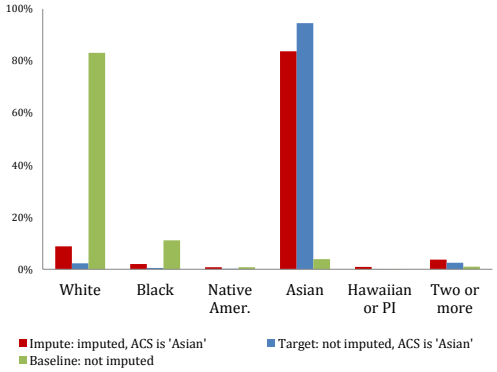
For the race categories with much smaller populations, the comparison to the ACS did not yield as accurate imputations. The groups Native American or Alaskan Native, and Hawaiian or Pacific Islander are 0.8% and 0.1%, respectively, of the U.S. population according to the 2000 Census. Conditional on having an ACS response in the same cate-



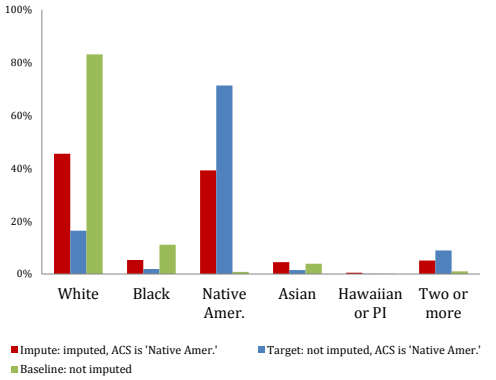
(a) White



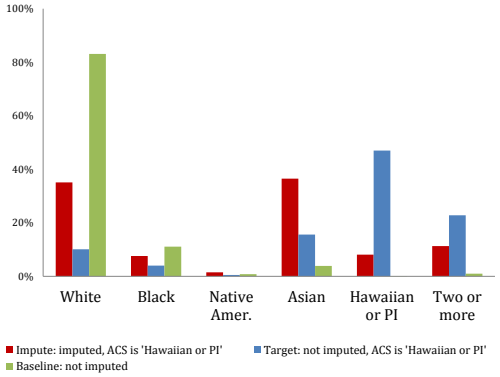
(b) Black



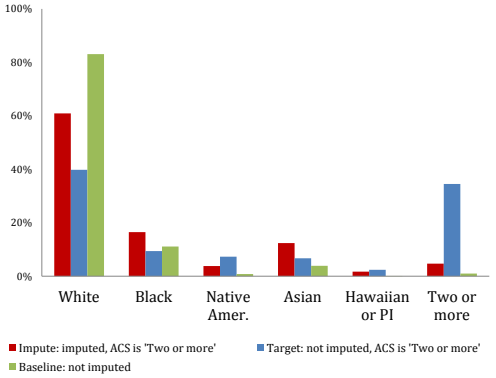
(c) Asian



(d) Native American or Alaska Native



(e) Hawaiian or Pacific Islander



(f) Two or More Races

Figure C.3: Impute versus Target: Race

Notes: Sub-figure titles correspond to the value of the race variable in the ICF. The blue bars show the distribution of race in the ICF among records that were not imputed. The red bars shows the distribution of race in the ICF among imputed records. The green bars show the distribution of race among records in the ICF that were not imputed *and* did not match to the ACS. The green bars do not vary across sub-figures. See Table C.3 for more detail.

Table C.2: Distribution of ICF Categories across ACS Response Categories, Ethnicity

Distribution of catagories in ICF	90% CI	Not Hispanic	Hispanic
Not in ACS			
Baseline: not imputed	Upper	90.7%	9.3%
	Mean	90.7%	9.3%
	Lower	90.7%	9.3%
ACS: Not Hispanic			
Impute: imputed, ACS is 'Not Hispanic'	Upper	94.7%	6.0%
	Mean	94.4%	5.6%
	Lower	94.0%	5.3%
Target: not imputed, ACS is 'Not Hispanic'	Upper	99.7%	0.4%
	Mean	99.6%	0.4%
	Lower	99.6%	0.3%
ACS: Hispanic			
Impute: imputed, ACS is 'Hispanic'	Upper	21.7%	81.7%
	Mean	20.0%	80.0%
	Lower	18.3%	78.3%
Target: not imputed, ACS is 'Hispanic'	Upper	5.5%	95.0%
	Mean	5.2%	94.8%
	Lower	5.0%	94.5%

Notes: 90% CI are 90% confidence intervals of the mean. Major row heading is the value of the ACS variable. Minor row heading is the value of the ICF variable. Major row header "Not in ACS" denotes records that did not match to the ACS.

gory, 39.2% were imputed into the Native American or Alaskan Native category (Figure C.3(d)), and 8.0% into Hawaiian or Pacific Islander (Figure C.3(e)). This is compared to target shares of 71.3% and 47.0%, respectively. For the Hawaiian or Pacific Islander, the majority of those responding as such on the ACS were imputed into the White and Asian categories at approximately equal rates. For Native American or Alaska Native, Figure C.3(d) shows the majority were imputed into the white category.

The category "Two or More Races" and "Some Other Race" also have inconsistent responses across input data. Those responding as "Two or More Races" are 1.0% of the population. Their target distribution is 34.5% of ACS respondents who report two or more races and who have the same response in the 2000 Census. For the records imputed from the 2000 Census who report two or more races in the ACS, only 4.7% were imputed into

the two or more races category. The other records were mostly imputed into the White, Black, and Asian categories as seen in Figure C.3(f). Note that "Some Other Race" is not an imputation category. Respondents to the ACS who answered "Some Other Race" were largely imputed to "White," with a large portion to "Two or More Races."

Comparison to the "D Sample"

The previous section examined the quality of the imputation at the person level. The next set of results asks how the imputation model fairs when used to reproduce LEHD public-use statistics. To do this, a simple comparison of key QWI variables is carried out for three sample states by education, and education x sex, using both reported education and imputed education. This analysis uses the "D sample," a sample of workers in the ICF that have a 2000 decennial long form education response. Here we compare beginning-of-quarter employment, full-quarter employment, and average quarterly wages for full-quarter employees using the QWI variables calculated using both respondent education and imputed education. The question of interest posed here is a simple one: for the sample of workers for whom reported education is known, do the QWI statistics show substantially different patterns when imputed education is used to tabulate the statistics rather than respondent education?

For this analysis beginning-of-quarter employment (B), full-quarter employment (F), and wages for full-quarter employees (Z_W3) are computed directly from the internal Employment History File, rather than the production system equivalent, using the standard definitions but not the fuzz factors. These indicators are computed for both the reported value of education and for each of the 10 implicates of the imputed education value. For the sake of simplicity in interpretation, we report the difference between the value of the indicator using reported education compared to the average value for the indicator using imputed education over the 10 implicates. While this is a simplification, as the variation

Table C.3: Distribution of ICF Categories across ACS Response Categories, Race

Distribution of catagories in ICF	90% CI	White	Black	Native Amer.	≥ Asian	Hawaiian & PI	≥ Two or More
Not in ACS							
Baseline: not imputed	Upper	83.1%	11.1%	0.8%	3.9%	0.1%	1.0%
	Mean	83.1%	11.1%	0.8%	3.9%	0.1%	1.0%
	Lower	83.1%	11.1%	0.8%	3.9%	0.1%	1.0%
ACS: White							
Impute: imputed, ACS is 'White'	Upper	94.8%	2.0%	1.2%	1.0%	0.4%	1.5%
	Mean	94.5%	1.9%	1.1%	0.9%	0.3%	1.4%
	Lower	94.2%	1.7%	0.9%	0.7%	0.3%	1.2%
Target: not imputed, ACS is 'White'	Upper	99.3%	0.2%	0.1%	0.1%	0.0%	0.4%
	Mean	99.3%	0.2%	0.1%	0.1%	0.0%	0.3%
	Lower	99.2%	0.2%	0.1%	0.1%	0.0%	0.3%
ACS: Black							
Impute: imputed, ACS is 'Black'	Upper	7.3%	90.8%	1.0%	1.2%	0.6%	3.0%
	Mean	6.3%	89.5%	0.6%	0.8%	0.4%	2.4%
	Lower	5.3%	88.3%	0.3%	0.4%	0.1%	1.7%
Target: not imputed, ACS is 'Black'	Upper	2.4%	96.9%	0.2%	0.2%	0.0%	0.9%
	Mean	2.2%	96.7%	0.1%	0.1%	0.0%	0.8%
	Lower	2.0%	96.5%	0.1%	0.1%	0.0%	0.7%
ACS: Native American							
Impute: imputed, ACS is 'Native Amer.'	Upper	52.9%	8.6%	46.5%	7.6%	1.6%	8.5%
	Mean	45.5%	5.3%	39.2%	4.5%	0.5%	5.1%
	Lower	38.0%	2.0%	31.8%	1.4%	0.0%	1.6%
Target: not imputed, ACS is 'Native Amer.'	Upper	18.1%	2.5%	73.4%	2.1%	0.3%	10.2%
	Mean	16.4%	1.9%	71.3%	1.5%	0.1%	8.9%
	Lower	14.7%	1.2%	69.2%	0.9%	0.0%	7.6%
ACS: Asian							
Impute: imputed, ACS is 'Asian'	Upper	10.7%	3.0%	1.5%	86.2%	1.6%	5.0%
	Mean	8.8%	2.0%	0.8%	83.7%	0.9%	3.7%
	Lower	6.9%	1.1%	0.2%	81.3%	0.3%	2.4%
Target: not imputed, ACS is 'Asian'	Upper	2.6%	0.6%	0.2%	94.9%	0.2%	2.8%
	Mean	2.3%	0.5%	0.2%	94.5%	0.1%	2.5%
	Lower	2.0%	0.3%	0.1%	94.0%	0.1%	2.1%
ACS: Hawaiian & PI							
Impute: imputed, ACS is 'Hawaiian & PI'	Upper	53.1%	17.5%	6.3%	54.7%	18.5%	23.3%
	Mean	35.1%	7.6%	1.5%	36.5%	8.1%	11.3%
	Lower	17.1%	0.0%	0.0%	18.3%	0.0%	0.0%
Target: not imputed, ACS is 'Hawaiian & PI'	Upper	13.7%	6.3%	1.4%	19.9%	53.0%	27.8%
	Mean	10.1%	4.0%	0.5%	15.6%	47.0%	22.8%
	Lower	6.5%	1.6%	0.0%	11.3%	41.1%	17.8%
ACS: Two or More							
Impute: imputed, ACS is 'Two or More'	Upper	89.4%	7.0%	3.3%	3.5%	1.0%	2.5%
	Mean	87.3%	5.6%	2.3%	2.5%	0.6%	1.7%
	Lower	85.2%	4.1%	1.4%	1.6%	0.1%	0.9%
Target: not imputed, ACS is 'Two or More'	Upper	85.1%	6.4%	2.9%	4.6%	0.4%	2.9%
	Mean	84.4%	5.9%	2.6%	4.2%	0.3%	2.6%
	Lower	83.6%	5.5%	2.2%	3.9%	0.2%	2.3%
ACS: Some Other							
Impute: imputed, ACS is 'Some Other'	Upper	64.8%	19.5%	5.3%	15.0%	2.7%	6.4%
	Mean	60.9%	16.5%	3.8%	12.4%	1.7%	4.7%
	Lower	57.0%	13.6%	2.3%	9.8%	0.6%	3.0%
Target: not imputed, ACS is 'Some Other'	Upper	41.0%	10.1%	8.0%	7.3%	2.8%	35.6%
	Mean	39.8%	9.4%	7.3%	6.7%	2.4%	34.5%
	Lower	38.6%	8.6%	6.7%	6.1%	2.0%	33.3%

Notes: 90% CI are 90% confidence intervals of the mean. Major row heading is the value of the ACS variable. Minor row heading is the value of the ICF variable. Major row header "Not in ACS" denotes records that did not match to the ACS.

over the implicates is typically small and generally much smaller than the difference between the average and reported values, it is consistent with the analysis done in the main text of the paper.

As can be seen in Table C.4, the comparisons at the state level generally show close correspondence between QWI values using reported education and imputed education, particularly for wages. At the state level, the difference between average full-quarter wages within education categories for reported and imputed education ranges from -6.7% to +8.0% with the smallest difference being less than 0.2%. The share of beginning-of-quarter employment in each education category varies by a range of -4.9 to 6.4 percentage points with the smallest difference being -0.1 percentage points at the statewide level. Overall, differences in the distribution of full-quarter employment between reported and imputed education are similar to those for beginning-of-quarter employment.

Table C.4: Comparison of QWI Variables for the Decennial Sample (D Sample): Actual vs. Imputed Education

	Employment Counts		B Employment Share		F Employment Share		Average Full-quarter wage, (Z_W3)	
Statewide Distribution	B	F	Actual	Imputed*	Actual	Imputed*	Actual	Imputed**
Delaware								
Less than High School	3,510	2,950	10.0%	10.1%	9.7%	9.6%	$6,100	$6,180
High School Graduate	11,400	9,910	32.7%	27.7%	32.4%	27.4%	$7,590	$7,590
Some College or Associates Degree	10,200	8,920	29.3%	30.6%	29.2%	30.6%	$8,950	$9,110
College Graduate or Greater	9,760	8,770	28.0%	31.5%	28.7%	32.4%	$14,900	$13,800
Illinois								
Less than High School	51,500	45,100	8.29%	9.16%	8.07%	8.90%	$6,390	$6,390
High School Graduate	168,000	151,000	27.00%	28.40%	27.00%	28.20%	$7,520	$7,730
Some College or Associates Degree	205,000	185,000	33.00%	31.40%	33.10%	31.50%	$8,880	$9,530
College Graduate or Greater	197,000	178,000	31.80%	31.00%	31.90%	31.40%	$15,700	$15,100
New Jersey								
Less than High School	31,500	27,300	9.14%	8.42%	8.93%	8.12%	$6,860	$6,510
High School Graduate	95,800	85,100	27.80%	21.40%	27.80%	21.10%	$8,550	$8,360
Some College or Associates Degree	93,100	82,600	27.00%	29.20%	27.00%	29.20%	$10,400	$10,500
College Graduate or Greater	125,000	111,000	36.10%	41.00%	36.30%	41.50%	$17,700	$16,400

Notes: *Average share over ten implicats. **Average over ten implicates. Statics computed for year 2000 quarter 2. B denotes beginning-of-quarter employment, and F denotes full-quarter employment.

For B, F, and Z_W3 for education x sex at the state level, the correspondence is again quite close. In Table C.5, the difference in education x sex tabulations between average full-quarter wages within categories for reported and imputed education ranges from -8.1% to +9.4% with the smallest difference being less than 0.09%. The share of beginning-

of-quarter employment in each education category varies by a range of -5.3 to 6.6 percentage points with the smallest difference being less than 0.001 percentage points at the statewide level. Differences in male/female wage gaps and employment by education across states are largely retained in the imputed results. Generally and not surprisingly, differences in state comparisons tend to be replicated in smaller cells as well. For instance for IL, the differences between B and F for imputed vs. reported education are very small at the state level and are also very small in the education x sex cells, while somewhat larger discrepancies in NJ and IL between some education categories are seen in education x sex cells for those two groups.

Table C.5: Comparison of QWI Variables for the Decennial Sample (D Sample) by Sex: Actual vs. Imputed Education

		Employment Counts		*B* Employment Share		*F* Employment Share		Average Full-quarter wage, (*Z_W3*)	
Statewide Distribution by Sex		*B*	*F*	Actual	Imputed*	Actual	Imputed*	Actual	Imputed**
Delaware									
Female	Less than High School	1,420	1,100	8.45%	8.59%	8.08%	8.04%	$4,470	$4,360
	High School Graduate	5,450	4,750	32.50%	27.10%	32.40%	26.80%	$5,810	$5,610
	Some College or Associates Degree	5,320	4,640	31.80%	32.40%	31.70%	32.40%	$7,120	$7,080
	College Graduate or Greater	4,570	4,080	27.20%	31.90%	27.90%	32.70%	$11,200	$10,600
Male	Less than High School	2,100	1,780	11.50%	11.50%	11.10%	11.00%	$7,190	$7,400
	High School Graduate	5,980	5,170	32.90%	28.30%	32.40%	28.00%	$9,220	$9,320
	Some College or Associates Degree	4,910	4,300	27.00%	29.00%	27.00%	29.00%	$10,900	$11,200
	College Graduate or Greater	5,300	4,700	28.60%	31.20%	29.50%	32.00%	$18,100	$16,900
Illinois									
Female	Less than High School	22,900	20,000	7.46%	8.35%	7.28%	8.11%	$4,500	$4,510
	High School Graduate	81,900	73,700	26.70%	28.90%	26.80%	28.80%	$5,400	$5,490
	Some College or Associates Degree	107,000	95,900	34.90%	33.30%	35.00%	33.40%	$6,600	$6,960
	College Graduate or Greater	95,000	84,900	31.00%	29.40%	30.90%	29.70%	$10,900	$10,800
Male	Less than High School	28,700	25,200	9.10%	9.96%	8.84%	9.66%	$7,880	$7,900
	High School Graduate	85,800	77,200	27.30%	28.00%	27.10%	27.70%	$9,550	$9,990
	Some College or Associates Degree	97,900	88,900	31.10%	29.60%	31.20%	29.60%	$11,300	$12,300
	College Graduate or Greater	102,000	93,400	32.50%	32.50%	32.80%	33.00%	$20,100	$18,900
New Jersey									
Female	Less than High School	14,000	12,100	8.15%	7.80%	7.96%	7.52%	$4,940	$4,730
	High School Graduate	49,300	43,900	28.70%	22.10%	28.80%	21.90%	$6,450	$6,070
	Some College or Associates Degree	49,200	43,600	28.60%	31.10%	28.70%	31.10%	$7,950	$7,890
	College Graduate or Greater	59,500	52,600	34.60%	39.00%	34.60%	39.50%	$12,700	$12,100
Male	Less than High School	17,500	15,200	10.10%	9.03%	9.90%	8.71%	$8,390	$8,040
	High School Graduate	46,500	41,200	26.90%	20.60%	26.80%	20.30%	$10,800	$10,800
	Some College or Associates Degree	43,900	38,900	25.40%	27.40%	25.30%	27.40%	$13,100	$13,600
	College Graduate or Greater	65,100	58,400	37.60%	42.90%	38.00%	43.50%	$22,100	$20,200

Notes: *Average share over ten implicats. **Average over ten implicates. Statics computed for year 2000 quarter 2. *B* denotes beginning-of-quarter employment, and *F* denotes full-quarter employment.

C.2 Imputation Procedure to Match Research Snapshot and Public-use Data

The research snapshot used to compute the total variance measures for the QWI differs from the production system used to create the public-use QWI files. The production system does not save the ten implicates to create the public-use QWI, but these implicates are necessary for the creation of the total variability measures. The research snapshot does not exactly replicate the production QWI statistics due to edits made to each snapshot, which are never reconciled. Due to these edits and rounding, it is sometimes the case that the computed statistics for a given cell do not exactly match. For cells with large employment counts this is a trivial concern as the variance for each statistic is already quite low, and small changes in the magnitude of the statistic result in marginal changes to the coefficient of variation. In cells with small employment counts (less than ten, say), this is not the case. Small changes in the size of the of employment count lead to large changes in the coefficient of variation. In this appendix we detail how we edit and scale the variance measures to account for the occasional differences in the internal and public-use statistics.

Before proceeding to the edit and scaling algorithm, a brief discussion of the reference distribution for the coefficient of variation is necessary. The intuition for the edit procedure is that our assumption of equivalent coefficient of variations for the public-use and research snapshots is "reasonable." For any cell with a given employment size, what is reasonable depends on the state, the demographic characteristics and the level of aggregation. We control for these confounding factors by performing the edit procedure separately for each state by ownership type by characteristic crossing. Next, we separate the data by beginning-of-quarter employment; full-quarter employment and average monthly full-quarter earnings; and flow employment and payroll. Within each of the three separate edits, we further separate each cell by its level of aggregation. The edit

algorithm is therefore run separately for each state by ownership type by characteristic crossing, by each of the three employment definitions governing the five statistics and by each level of aggregation.

After partitioning the data, the edit algorithm then proceeds as follows. First, we calculate one percent quantiles of the internally calculated employment statistic from the minimum to the maximum. We collapse bins where the employment count is the same for consecutive quantiles leaving us with at most 100 bins for the internally calculated employment statistic. For each bin we calculate the 5th and 95th percentile of the coefficient of variation for the employment statistic as well as average monthly earnings and payroll for full-quarter employment and total employment, respectively. In addition, for each of the five QWI statistics we calculate the median within and between variance as well as the median statistic in the bin.

Once the bins are set, for each record we look-up the bin associated with each of the three public-use employment statistics. If the coefficient of variation for the internally calculated statistic falls either below the 5th percentile or above the 95th percentile of the coefficient of variation in that bin, we use the median within-variance and the median between-variance for that statistic and rescale them accordingly. We then make the total variance, missingness ratio, and degrees of freedom calculations from our edited within- and between-variance. An example will elucidate the procedure.

Suppose we have a cell with an internally calculated flow employment (M) count of 5 and due to edits and rounding the public-use statistic ($EmpTotal$)is 7. As is typical for low-levels of aggregation and small employment counts, the bins consist of only cells with the same employment counts. That is, the bins consist only of cells with counts of 5, 6, 7, etc. Our public-use flow employment total is 7, so we look at the bin of cells with flow employment counts of 7 and compare our internally calculated coefficient of variation to the distribution in that bin. The coefficient of variation for this cell was calculated from

an internal count of 5 and the coefficient of variation is in this example greater than the 95th percentile in the cell. We therefore assign the public-use statistic the median within- and between-variance from the bin, and we scale the two variances by the median flow employment count in the bin, which in this example is simply 7. In this example the median flow employment count in the bin is the same as the public-use statistic negating any change in the variance from scaling, but we use a more reasonable estimate of the within- and between-variance.

C.3 Handling Structural and Sampling Zeros

The public-use QWI files are sparse. If a given cell does not have at least one dollar from a UI-covered job, the cell does not appear in the released data. However, just because a cell does not appear in a particular quarter does not mean that it will not appear in a subsequent quarter. If a cell contains zeros given quarter, for some combinations of stratifiers, but not others, then there are firms operating in that cell, and the zeros are sampling zeros. If there no evidence of any firm activity in that cell–meaning all combinations of stratifiers show zero employment, then those zeros are all structural zeros. We supplement the unemployment insurance records used as the core inputs to the QWI with firm reports from the QCEW. The QCEW are a firm-level virtual-census of employment and wages comprising the universe of firms covered by state unemployment insurance systems and some federal employment. The universe of firm activity in the QWI and the QCEW is quite similar but it does not perfectly overlap. To infer firm activity in a given state, year, quarter, county, and NAICS Sector cell, which is the correct frame for distinguishing sampling from structural zeros, we use the union of firm activity from the QWI and QCEW universes. If a cell does not appear in the unemployment insurance microdata, but we find evidence of firm activity – any positive employment in any month or

positive wages – from the QCEW we add that cell to the public use file, including all lower levels of aggregation. We flag all sampling zeros with the variable "sample_zero." The five QWI statistics for all sampling zeros are set to zero, and we impute each of their variability statistics.

We impute the variability measures for sampling zeros by exploiting the edit procedure in Appendix C.2. Recall that in the edit procedure we calculate various moments of the coefficient of variation, within-variance, and between-variance distributions by bins of the internally calculated employment size. The bins are calculated separately for each state, ownership type, characteristic crossing, and aggregation level. We use the median within- and between-variance from the zero bin as the sample zero within- and between-variance. In cases where the aggregation level is too high so as no zero bin exists, we drop down to the next lowest level of aggregation where a zero bin is available and calculate the ratio of the coefficient of variation for the one and zero bins. We scale the within- and between-variance at our reference level of aggregation using the one bin and the ratio calculated from the lower level of aggregation. To summarize, the median within- and between- variance from the zero bin of the edit procedure are used as our imputation of the within-and between variance for sampling zeros. We then derive the total variance, missingness ratio, and degrees of freedom estimates from the within- and between-variance.

C.4 Data Notes

- North Carolina, Colorado, and Massachusetts are not in the R2012Q4 QWI release and have not been included in the variability files.
- 720 records from the Georgia age by sex all employment file, 588 records from the Georgia race by ethnicity all employment file, and 420 records from the Georgia sex

by education all employment file include the NAICS sector 99. This is an error in the release, and these records have been removed from their respective variability files.

Table C.6: Summary of Total Variability of Private Total Employment (*EmpTotal*) by Table and Count

Table and *EmpTotal* count range	Proportion of Cells	Number of Cells	Median Count	Median Total Variation	Median Rubin Missingness Rate (Percent)	Quantiles of Coefficient of Variation			Median Approximate 90% Confidence Intervals Margin of Error		
						5th	Median	95th	Median df	Count	Percent
Private											
Age x Gender											
+1000	1.0000	46,480	80,832	8250.00	43.60%	0.0004	0.0011	0.0036	47	118	0.14%
Race x Ethnicity											
10-99	0.0199	695	53	49.80	96.10%	0.0836	0.1409	0.2605	9	10	19.48%
100-999	0.1306	4,553	452	411.00	95.30%	0.0242	0.0469	0.0935	9	28	6.49%
+1000	0.8495	29,612	13,614	5970.00	86.40%	0.0002	0.0044	0.0274	12	105	0.59%
Gender x Education											
+1000	1.0000	23,240	157,493	192000.00	96.60%	0.0013	0.0031	0.0086	9	606	0.43%
Industry x County											
zero measured value, after rounding	0.0045	12,955	0	0.30	94.40%	(a)	(a)	(a)	10	1	(a)
1-2	0.0001	188	1	0.42	79.35%	0.1982	0.3913	0.9553	14	1	52.73%
3-9	0.0158	45,741	7	0.50	0.00%	0.0614	0.1073	0.3800	9999	1	13.76%
10-99	0.2603	753,131	46	4.82	15.40%	0.0242	0.0511	0.1619	380	3	6.56%
100-999	0.4402	1,273,670	294	53.70	66.30%	0.0105	0.0235	0.0591	20	10	3.12%
+1000	0.2792	807,706	3,058	822.00	76.60%	0.0023	0.0081	0.0201	15	38	1.09%
Age x Gender x Industry x County											
zero measured value, after rounding	0.2019	8,888,449	0	0.21	95.10%	(a)	(a)	(a)	9	1	(a)
1-2	0.0050	220,862	2	0.35	67.20%	0.1564	0.3066	0.7280	19	1	40.71%
3-9	0.2171	9,557,988	5	0.81	61.50%	0.0885	0.1722	0.3887	23	1	22.73%
10-99	0.3796	16,713,425	27	5.35	69.70%	0.0382	0.0815	0.1800	18	3	10.84%
100-999	0.1622	7,142,409	225	55.20	75.40%	0.0142	0.0303	0.0618	15	10	4.06%
+1000	0.0343	1,511,324	1,972	502.00	75.70%	0.0041	0.0103	0.0201	15	30	1.38%
Race x Ethnicity x Industry x County											
zero measured value, after rounding	0.5839	19,047,330	0	0.20	95.20%	(a)	(a)	(a)	9	1	(a)
1-2	0.0058	190,628	2	0.69	91.70%	0.2636	0.6229	0.9257	10	1	85.47%
3-9	0.1330	4,339,494	5	2.35	88.90%	0.1325	0.3167	0.5963	11	2	43.18%
10-99	0.1607	5,240,825	26	10.10	85.30%	0.0426	0.1162	0.2704	12	4	15.76%
100-999	0.0830	2,707,617	249	75.90	79.90%	0.0137	0.0322	0.0745	14	12	4.33%
+1000	0.0336	1,094,612	2,586	766.00	79.20%	0.0031	0.0095	0.0212	14	37	1.28%
Gender x Education x Industry x County											
zero measured value, after rounding	0.0996	2,207,640	0	0.26	94.80%	(a)	(a)	(a)	10	1	(a)
1-2	0.0050	111,105	2	1.35	93.10%	0.4281	0.6538	0.9466	10	2	89.72%
3-9	0.2024	4,484,091	5	3.99	92.70%	0.2395	0.3791	0.6106	10	3	52.03%
10-99	0.4264	9,446,881	29	22.00	92.80%	0.0865	0.1624	0.2961	10	6	22.28%
100-999	0.2119	4,695,684	235	190.00	93.10%	0.0292	0.0569	0.0967	10	19	7.81%
+1000	0.0546	1,209,869	2,089	1790.00	93.50%	0.0087	0.0193	0.0321	10	58	2.65%

Notes: Total employment is defined as all jobs held by a worker at the same establishment during the quarter. Statistics are computed across all state-year-quarters within a table. The "Private" category of establishments includes only private establishments. All tables include all valid QWI age groups with the exception of any table including education, in which case only jobs with workers age 25 and older are included. For statistic definitions for total employment, please see their respective equations in the accompanying text: Count 3.6, Total Variation 3.15, Missingness Ratio 3.16, Coefficient of Variation 3.32. (a) Undefined value.

Table C.7: Summary of Total Variability of Private Beginning-of-Quarter Employment (*Emp*) by Table and Count

Table and *Emp* count range	Proportion of Cells	Number of Cells	Median Count	Median Total Variation	Median Rubin Missingness Rate (Percent)	Quantiles of Coefficient of Variation			Median Approximate 90% Confidence Intervals Margin of Error		
						5th	Median	95th	Median df	Count	Percent
Private											
Age x Gender											
+1000	1.0000	45,712	61,308	5000.00	37.50%	0.0003	0.0011	0.0035	64	92	0.14%
Race x Ethnicity											
3-9	0.0000	1	9	3.57	92.60%	0.2099	0.2099	0.2099	10	3	28.81%
10-99	0.0282	968	47	36.80	95.90%	0.0783	0.1282	0.2729	9	8	17.73%
100-999	0.1607	5,509	466	328.00	94.70%	0.0115	0.0427	0.0828	10	25	5.86%
+1000	0.8111	27,806	11,716	4000.00	83.20%	0.0002	0.0041	0.0241	13	85	0.55%
Gender x Education											
+1000	1.0000	22,856	132,437	136000.00	96.60%	0.0013	0.0031	0.0085	9	510	0.42%
Industry x County											
zero measured value, after rounding	0.0096	27,314	0	0.29	95.50%	(a)	(a)	(a)	9	1	(a)
1-2	0.0001	313	2	0.37	66.90%	0.1425	0.3279	0.9000	20	1	43.45%
3-9	0.0242	69,274	7	0.42	0.00%	0.0561	0.1009	0.3648	9999	1	12.93%
10-99	0.2916	833,370	44	4.28	21.20%	0.0227	0.0499	0.1609	201	3	6.42%
100-999	0.4286	1,225,043	283	51.40	71.10%	0.0102	0.0235	0.0586	17	10	3.14%
+1000	0.2459	702,856	2,936	747.00	78.70%	0.0024	0.0080	0.0199	14	37	1.08%
Age x Gender x Industry x County											
zero measured value, after rounding	0.2368	10,146,295	0	0.20	95.60%	(a)	(a)	(a)	9	1	(a)
1-2	0.0051	217,085	2	0.33	69.70%	0.1409	0.2971	0.7099	18	1	39.52%
3-9	0.2243	9,610,779	5	0.72	63.10%	0.0811	0.1641	0.3815	22	1	21.69%
10-99	0.3624	15,526,526	26	4.98	72.80%	0.0365	0.0797	0.1786	17	3	10.63%
100-999	0.1432	6,136,088	222	51.50	77.80%	0.0135	0.0296	0.0610	14	10	3.97%
+1000	0.0282	1,210,018	1,931	452.00	77.40%	0.0042	0.0101	0.0196	15	29	1.35%
Race x Ethnicity x Industry x County											
zero measured value, after rounding	0.6229	20,152,114	0	0.19	95.70%	(a)	(a)	(a)	9	1	(a)
1-2	0.0052	168,667	2	0.66	92.30%	0.2579	0.6042	0.8972	10	1	82.90%
3-9	0.1222	3,951,621	5	2.16	89.60%	0.1241	0.3040	0.5810	11	2	41.45%
10-99	0.1465	4,740,254	26	9.16	85.80%	0.0395	0.1103	0.2619	12	4	14.96%
100-999	0.0746	2,411,633	246	69.90	81.40%	0.0130	0.0310	0.0716	13	11	4.18%
+1000	0.0287	926,867	2,513	687.00	80.90%	0.0031	0.0093	0.0205	13	35	1.25%
Gender x Education x Industry x County											
zero measured value, after rounding	0.1166	2,517,116	0	0.26	95.40%	(a)	(a)	(a)	9	1	(a)
1-2	0.0055	118,679	2	1.34	93.80%	0.4257	0.6500	0.9392	10	2	89.19%
3-9	0.2150	4,638,908	5	3.87	93.50%	0.2365	0.3763	0.6062	10	3	51.64%
10-99	0.4195	9,052,346	28	21.00	93.60%	0.0857	0.1620	0.2946	10	6	22.23%
100-999	0.1959	4,228,095	232	183.00	93.90%	0.0288	0.0563	0.0957	10	19	7.73%
+1000	0.0475	1,025,888	2,045	1670.00	94.20%	0.0086	0.0191	0.0315	10	56	2.62%

Notes: Beginning-of-quarter employment is defined as all jobs held by a worker at the same establishment during the quarter and during the previous quarter. Statistics are computed across all state-year-quarters within a table. The "Private" category of establishments includes only private establishments. All tables include all valid QWI age groups with the exception of any table including education, in which case only jobs with workers age 25 and older are included. For statistic definitions for beginning of quarter employment, please see their respective equations in the accompanying text: Count 3.6, Total Variation 3.15, Missingness Ratio 3.16, Coefficient of Variation 3.32. (a) Undefined value.

Table C.8: Summary of Total Variability of Private Full-Quarter Employment (*EmpS*) by Table and Count

Table and *EmpS* count range	Proportion of Cells	Number of Cells	Median Count	Median Total Variation	Median Rubin Missingness Rate (Percent)	Quantiles of Coefficient of Variation			Median Approximate 90% Confidence Intervals Margin of Error		
						5th	Median	95th	Median df	Count	Percent
Private											
Age x Gender											
100-999	0.0001	6	965	414.00	79.70%	0.0211	0.0211	0.0211	14	27	2.84%
+1000	0.9999	44,938	50,251	3810.00	33.10%	0.0004	0.0012	0.0038	82	80	0.15%
Race x Ethnicity											
3-9	0.0005	17	9	8.14	96.60%	0.1746	0.3394	0.4180	9	4	46.94%
10-99	0.0351	1,184	46	32.70	95.10%	0.0747	0.1279	0.2935	9	8	17.69%
100-999	0.1780	6,001	452	301.00	94.40%	0.0133	0.0420	0.0856	10	24	5.76%
+1000	0.7863	26,506	10,312	3290.00	80.60%	0.0002	0.0042	0.0239	13	77	0.56%
Gender x Education											
+1000	1.0000	22,472	115,661	114000.00	96.40%	0.0014	0.0031	0.0088	9	467	0.43%
Industry x County											
zero measured value, after rounding	0.0142	40,036	0	0.28	95.60%	(a)	(a)	(a)	9	1	(a)
1-2	0.0002	505	2	0.19	0.00%	0.1308	0.2518	0.8485	9999	1	32.27%
3-9	0.0327	92,014	7	0.40	0.00%	0.0571	0.1024	0.3636	9999	1	13.12%
10-99	0.3147	886,839	43	4.21	23.50%	0.0231	0.0509	0.1608	162	3	6.55%
100-999	0.4159	1,172,234	276	51.60	71.80%	0.0105	0.0241	0.0589	17	10	3.22%
+1000	0.2224	626,901	2,870	723.00	78.40%	0.0024	0.0081	0.0200	14	36	1.10%
Age x Gender x Industry x County											
zero measured value, after rounding	0.2674	11,179,951	0	0.20	95.60%	(a)	(a)	(a)	9	1	(a)
1-2	0.0052	216,857	2	0.33	69.00%	0.1409	0.2958	0.7036	18	1	39.35%
3-9	0.2274	9,509,759	5	0.71	62.50%	0.0808	0.1639	0.3791	23	1	21.63%
10-99	0.3462	14,475,571	26	4.93	72.70%	0.0367	0.0805	0.1797	17	3	10.73%
100-999	0.1295	5,413,281	221	50.70	77.40%	0.0134	0.0295	0.0611	15	10	3.96%
+1000	0.0243	1,017,595	1,896	429.00	76.50%	0.0041	0.0099	0.0194	15	28	1.33%
Race x Ethnicity x Inudstry x County											
zero measured value, after rounding	0.6503	20,835,268	0	0.19	95.80%	(a)	(a)	(a)	9	1	(a)
1-2	0.0049	157,195	2	0.65	92.20%	0.2579	0.6021	0.8874	10	1	82.62%
3-9	0.1144	3,664,172	5	2.08	89.10%	0.1213	0.2990	0.5761	11	2	40.76%
10-99	0.1366	4,375,170	26	8.75	84.90%	0.0385	0.1074	0.2582	12	4	14.56%
100-999	0.0685	2,195,777	244	67.70	80.80%	0.0131	0.0306	0.0705	13	11	4.14%
+1000	0.0253	810,388	2,467	663.00	80.30%	0.0031	0.0093	0.0204	13	35	1.26%
Gender x Education x Industry x County											
zero measured value, after rounding	0.1317	2,774,036	0	0.26	95.30%	(a)	(a)	(a)	9	1	(a)
1-2	0.0059	124,663	2	1.32	93.80%	0.4278	0.6500	0.9368	10	2	89.19%
3-9	0.2237	4,709,531	5	3.83	93.50%	0.2365	0.3766	0.6065	10	3	51.68%
10-99	0.4122	8,679,900	27	20.60	93.60%	0.0860	0.1633	0.2954	10	6	22.40%
100-999	0.1839	3,871,290	230	181.00	93.80%	0.0288	0.0564	0.0958	10	18	7.74%
+1000	0.0426	897,250	2,011	1640.00	94.20%	0.0086	0.0192	0.0315	10	56	2.64%

Notes: Total employment is defined as all jobs held by a worker at the same establishment during the quarter. Statistics are computed across all state-year-quarters within a table. The "Private" category of establishments includes only private establishments. All tables include all valid QWI age groups with the exception of any table including education, in which case only jobs with workers age 25 and older are included. For statistic definitions for total employment, please see their respective equations in the accompanying text: Count 3.6, Total Variation 3.15, Missingness Ratio 3.16, Coefficient of Variation 3.32. (a) Undefined value.

Table C.9: Summary of Total Variability of Private Total Payroll (*Payroll*) by Table and Count

Table and *EmpTotal* count range	Proportion of Cells	Number of Cells	Median Payroll	Median Total Variation	Median Rubin Missingness Rate (Percent)	Quantiles of Coefficient of Variation			Median Approximate 90% Confidence Intervals Margin of Error		
						5th	Median	95th	Median df	Count	Percent
Private											
Age x Gender											
+1000	1.0000	46,480	375,627,224.50	3.96E+11	29.50%	0.0005	0.0016	0.0083	104	811,617.46	0.21%
Race x Ethnicity											
10-99	0.0199	695	233,792.00	2.29E+09	97.30%	0.1187	0.2083	0.4509	9	66,183.38	28.80%
100-999	0.1306	4,553	2,166,851.00	1.85E+10	96.10%	0.0334	0.0678	0.1493	9	188,112.25	9.38%
+1000	0.8495	29,612	71,561,892.50	4.67E+11	82.40%	0.0005	0.0070	0.0479	13	922,671.93	0.94%
Gender x Education											
+1000	1.0000	23,240	1,122,441,816.50	2.05E+13	96.10%	0.0018	0.0042	0.0117	9	6,261,928.94	0.57%
Industry x County											
zero measured value, after rounding	0.0040	12,955	0.00	9.45E+06	99.80%	0.0392	0.4429	1.0808	9	4,251.55	61.25%
1-2	0.1144	373,579	39,363.00	8.17E+06	0.00%	0.0000	0.0671	0.5900	9999	3,663.33	8.60%
3-9	0.0140	45,741	28,035.00	5.56E+06	0.00%	0.0302	0.0834	0.5458	9999	3,022.05	10.68%
10-99	0.2305	753,131	214,551.00	1.23E+08	9.57%	0.0198	0.0536	0.2327	984	14,222.64	6.88%
100-999	0.3899	1,273,670	1,568,970.00	2.55E+09	76.50%	0.0108	0.0307	0.0905	15	67,697.26	4.12%
+1000	0.2473	807,706	19,017,146.50	6.38E+10	80.70%	0.0035	0.0118	0.0329	13	341,035.20	1.59%
Age x Gender x Industry x County											
zero measured value, after rounding	0.1701	8,888,449	0.00	6.05E+05	100.00%	0.0000	0.2197	1.3084	9	1,075.74	30.39%
1-2	0.1618	8,454,917	4,523.00	3.34E+05	12.30%	0.0000	0.1107	0.8974	598	741.46	14.20%
3-9	0.1829	9,557,988	17,743.00	8.87E+06	83.20%	0.0453	0.1636	0.6023	13	4,021.15	22.08%
10-99	0.3198	16,713,425	113,931.00	1.16E+08	85.80%	0.0327	0.0939	0.2689	12	14,606.91	12.73%
100-999	0.1367	7,142,409	1,166,690.00	2.00E+09	87.20%	0.0144	0.0384	0.0942	11	60,974.46	5.23%
+1000	0.0289	1,511,324	14,164,728.00	3.60E+10	84.30%	0.0049	0.0135	0.0327	12	257,324.15	1.83%
Race x Ethnicity x Inudstry x County											
zero measured value, after rounding	0.4859	19,047,330	0.00	3.28E+06	100.00%	0.0115	0.4685	1.5608	9	2,504.77	64.80%
1-2	0.1727	6,771,506	4,934.00	6.96E+06	98.50%	0.0745	0.5842	1.2527	9	3,648.68	80.80%
3-9	0.1107	4,339,494	20,418.00	6.02E+07	96.70%	0.1040	0.4024	0.8577	9	10,730.73	55.65%
10-99	0.1337	5,240,825	125,020.00	3.38E+08	93.50%	0.0423	0.1488	0.3976	10	25,227.29	20.42%
100-999	0.0691	2,707,617	1,330,383.00	3.50E+09	88.60%	0.0153	0.0433	0.1145	11	80,661.63	5.90%
+1000	0.0279	1,094,612	16,504,365.00	5.75E+10	84.60%	0.0044	0.0135	0.0349	12	325,209.49	1.83%
Gender x Education x Industry x County											
zero measured value, after rounding	0.0845	2,207,640	153.00	2.35E+05	99.50%	0.0000	0.1886	2.1637	9	670.45	26.08%
1-2	0.1565	4,089,395	6,804.00	1.91E+07	98.80%	0.3127	0.6056	1.1470	9	6,044.33	83.76%
3-9	0.1716	4,484,091	25,232.00	1.38E+08	97.90%	0.2627	0.4790	0.8323	9	16,246.91	66.24%
10-99	0.3615	9,446,881	160,107.00	1.10E+09	97.20%	0.1026	0.2084	0.4189	9	45,869.87	28.83%
100-999	0.1797	4,695,684	1,522,171.00	1.37E+10	96.90%	0.0367	0.0763	0.1454	9	161,879.36	10.55%
+1000	0.0463	1,209,869	17,075,678.00	2.34E+11	96.40%	0.0118	0.0277	0.0559	9	669,020.05	3.83%

Notes: Total Payroll is defined only over total employment. It is calculated by summing the earnings for the reference quarter for total employment. See the table on total employment for the relevent counts. Statistics are computed across all state-year-quarters within a table. The "Private" category of establishments includes private only private establishments. All tables include all valid QWI age groups with the exception of any table including education, in which case only jobs with workers age 25 and older are included. For statistic definitions for beginning of quarter employment, please see their respective equations in the accompanying text: Total payroll 3.28, Total Variation 3.31, Missingness Ratio 3.16, Coefficient of Variation 3.32. (a) Undefined value.

Table C.10: Summary of Total Variability of Private Average Monthly Earnings (*EarnS*) by Table and Count

Table and *EmpS* count range	Proportion of Cells	Number of Cells	Median Average Monthly Earnings	Median Total Variation	Median Rubin Missingness Rate (Percent)	Quantiles of Coefficient of Variation			Median Approximate 90% Confidence Intervals Margin of Error		
						5th	Median	95th	Median df	Count	Percent
Private											
Age x Gender											
100-999	0.0001	6	1,686.00	13,600.00	87.00%	0.0691	0.0691	0.0691	11	159.00	9.42%
+1000	0.9999	44,938	2,146.00	6.79	22.90%	0.0004	0.0013	0.0066	171	3.35	0.17%
Race x Ethnicity											
3-9	0.0005	17	2,409.00	361,000.00	96.80%	0.1451	0.2600	0.6976	9	830.97	35.95%
10-99	0.0351	1,184	2,106.50	71,000.00	95.50%	0.0605	0.1252	0.3384	9	368.52	17.31%
100-999	0.1780	6,001	2,189.00	8,490.00	94.50%	0.0147	0.0425	0.1042	10	126.43	5.84%
+1000	0.7863	26,506	2,471.00	168.00	73.70%	0.0004	0.0052	0.0321	16	17.33	0.69%
Gender x Education											
+1000	1.0000	22,472	2,847.00	84.60	94.40%	0.0013	0.0032	0.0088	10	12.62	0.44%
Industry x County											
zero measured value, after rounding	0.0020	6,177	0.00	2,140,000.00	99.30%	(a)	(a)	(a)	9	2023.20	(a)
1-2	0.1096	342,768	2,088.00	7,230.00	0.00%	0.0000	0.0545	0.2903	9999	108.98	6.98%
3-9	0.0294	92,014	1,523.00	8,540.00	0.00%	0.0202	0.0655	0.2728	9999	118.44	8.39%
10-99	0.2836	886,839	1,957.00	5,240.00	11.20%	0.0150	0.0385	0.1256	714	92.85	4.93%
100-999	0.3749	1,172,234	2,265.00	2,020.00	65.80%	0.0081	0.0207	0.0551	20	59.57	2.75%
+1000	0.2005	626,901	2,701.00	414.00	70.60%	0.0026	0.0080	0.0216	18	27.07	1.07%
Age x Gender x Industry x County											
zero measured value, after rounding	0.0025	98,832	0.00	2,690,000.00	99.60%	(a)	(a)	(a)	9	2268.34	(a)
1-2	0.2269	8,953,462	1,297.00	9,140.00	0.00%	0.0000	0.0850	0.4718	9999	122.53	10.90%
3-9	0.2410	9,509,759	1,479.00	17,400.00	66.90%	0.0280	0.0932	0.2942	20	174.82	12.35%
10-99	0.3668	14,475,571	1,828.00	8,620.00	74.70%	0.0210	0.0538	0.1433	16	124.11	7.19%
100-999	0.1372	5,413,281	2,300.00	2,280.00	77.30%	0.0087	0.0224	0.0545	15	64.01	3.00%
+1000	0.0258	1,017,595	3,109.00	578.00	72.90%	0.0033	0.0084	0.0204	16	32.14	1.12%
Race x Ethnicity x Inudstry x County											
zero measured value, after rounding	0.0043	74,124	0.00	6,290,000.00	99.90%	(a)	(a)	(a)	9	3468.62	(a)
1-2	0.3501	5,991,260	1,835.00	226,000.00	97.20%	0.0499	0.2686	0.7331	9	657.48	37.15%
3-9	0.2141	3,664,172	1,942.00	126,000.00	93.50%	0.0563	0.1853	0.4744	10	487.08	25.42%
10-99	0.2557	4,375,170	2,082.00	24,500.00	86.80%	0.0246	0.0757	0.2127	11	213.41	10.33%
100-999	0.1283	2,195,777	2,290.00	3,170.00	79.90%	0.0097	0.0253	0.0661	14	75.73	3.40%
+1000	0.0474	810,388	2,723.00	508.00	75.50%	0.0032	0.0087	0.0221	15	30.22	1.16%
Gender x Education x Industry x County											
zero measured value, after rounding	0.0037	83,776	0.00	3,250,000.00	98.60%	(a)	(a)	(a)	9	2493.29	(a)
1-2	0.1917	4,326,849	1,800.00	424,000.00	98.10%	0.1419	0.3679	0.8320	9	900.56	50.88%
3-9	0.2087	4,709,531	1,908.00	241,000.00	95.90%	0.1254	0.2615	0.5459	9	678.95	36.16%
10-99	0.3846	8,679,900	2,210.00	63,700.00	94.40%	0.0534	0.1158	0.2588	10	346.32	15.89%
100-999	0.1715	3,871,290	2,558.00	12,000.00	94.30%	0.0203	0.0440	0.0937	10	150.32	6.04%
+1000	0.0398	897,250	3,175.00	2,870.00	94.20%	0.0076	0.0174	0.0408	10	73.51	2.38%

Notes: Average Monthly Earnings is defined only over full-quarter jobs. It is calculated by taking the earnings for the reference quarter for full-quarter jobs and dividing by 3. See the table on full-quarter employment for the relevent counts. Statistics are computed across all state-year-quarters within a table. The "Private" category of establishments includes private only private establishments.. All tables include all valid QWI age groups with the exception of any table including education, in which case only jobs with workers age 25 and older are included. For statistic definitions for beginning of quarter employment, please see their respective equations in the accompanying text: Average Monthly Earnings 3.20, Total Variation 3.23, Missingness Ratio 3.16, Coefficient of Variation 3.32. (a) Undefined value.

Table C.11: Between Variance of Beginning-of-Quarter (B) Population Counts

		Coefficient of Variation			
	Cell Count	Mean	Std Dev	Minimum	Maximum
A: Establishment Type and Age Range					
Population					
All Valid QWI Ages, All Establishments	2,957	1.059E-05	6.175E-06	1.738E-06	5.334E-05
All Valid QWI Ages, Private Establishments	2,957	1.187E-05	6.911E-06	1.984E-06	6.670E-05
B: State					
Postal Code					
AK	188	6.521E-05	5.021E-05	6.874E-06	2.234E-04
AL	172	3.040E-05	2.319E-05	3.695E-06	8.993E-05
AR	148	3.248E-05	2.201E-05	7.540E-06	7.745E-05
AZ	124	4.091E-05	3.357E-05	6.086E-06	1.385E-04
CA	324	2.463E-05	2.074E-05	2.581E-06	9.171E-05
CT	252	3.362E-05	2.643E-05	5.869E-06	1.417E-04
DC	104	8.532E-05	5.890E-05	1.611E-05	1.984E-04
DE	212	6.939E-05	5.802E-05	8.776E-06	2.319E-04
FL	304	1.812E-05	1.435E-05	2.625E-06	7.097E-05
GA	220	3.069E-05	2.497E-05	3.249E-06	1.027E-04
HI	256	4.458E-05	4.397E-05	5.854E-06	2.369E-04
IA	208	2.941E-05	2.218E-05	4.336E-06	1.009E-04
ID	332	6.875E-05	5.681E-05	6.696E-06	3.292E-04
IL	348	2.606E-05	2.101E-05	2.660E-06	7.849E-05
IN	220	2.383E-05	1.818E-05	2.140E-06	6.173E-05
KS	300	3.931E-05	3.122E-05	5.063E-06	1.262E-04
KY	172	2.659E-05	1.998E-05	3.718E-06	7.896E-05
LA	268	2.138E-05	1.482E-05	4.411E-06	6.201E-05
MD	348	3.166E-05	2.787E-05	3.609E-06	1.276E-04
ME	248	3.069E-05	2.227E-05	5.408E-06	9.439E-05
MI	180	1.896E-05	1.498E-05	3.122E-06	5.559E-05
MN	276	2.373E-05	1.933E-05	3.230E-06	6.993E-05
MO	268	2.291E-05	1.892E-05	2.885E-06	6.340E-05
MS	132	3.457E-05	2.448E-05	5.182E-06	8.206E-05
MT	300	4.375E-05	3.270E-05	7.252E-06	1.399E-04
ND	220	4.782E-05	3.732E-05	6.366E-06	1.632E-04
NE	204	3.910E-05	3.037E-05	6.833E-06	1.090E-04
NH	140	4.042E-05	2.809E-05	7.394E-06	9.946E-05
NJ	252	2.982E-05	2.341E-05	3.494E-06	1.275E-04
NM	260	7.141E-05	6.564E-05	6.204E-06	3.728E-04
NV	220	6.333E-05	5.146E-05	6.337E-06	1.739E-04
NY	188	2.334E-05	2.095E-05	2.948E-06	1.143E-04
OH	188	1.475E-05	1.147E-05	2.241E-06	4.309E-05
OK	188	4.435E-05	3.440E-05	4.895E-06	1.131E-04
OR	332	3.848E-05	2.973E-05	5.838E-06	1.269E-04
PA	236	1.181E-05	8.594E-06	1.738E-06	3.660E-05
RI	268	6.231E-05	4.507E-05	6.479E-06	1.753E-04
SC	220	3.604E-05	2.772E-05	4.599E-06	9.803E-05
SD	220	5.157E-05	4.009E-05	6.421E-06	1.497E-04
TN	220	2.394E-05	1.913E-05	2.935E-06	8.213E-05
TX	268	1.945E-05	1.583E-05	2.284E-06	7.985E-05
UT	196	6.618E-05	5.351E-05	7.202E-06	1.792E-04
VA	220	2.814E-05	2.341E-05	4.003E-06	1.111E-04
VT	188	4.439E-05	3.280E-05	5.941E-06	1.317E-04
WA	348	3.202E-05	2.560E-05	4.074E-06	9.715E-05
WI	348	2.128E-05	1.751E-05	1.960E-06	7.021E-05
WV	236	2.298E-05	1.554E-05	3.872E-06	7.009E-05
WY	172	8.874E-05	6.879E-05	1.586E-05	2.701E-04

Notes: There is small amount of between-implicate variance of state counts for beginning-of-quarter employment. We summarize the between variance using the coefficient of variation defined as the square root of the between-implicate variance divided by the average between-implicate weighted counts. Panel A summarizes the coefficient of variation for the between variance for the four different types of ownership type and age populations. The summary is taken across all state-year-quarters. Panel B summarizes the coefficient of variation for all states across all year, quarters, and ownership types and age range combinations.

Table C.12: Between Variance of Full-Quarter (F) Population Counts

		Coefficient of Variation			
	Cell Count	Mean	Std Dev	Minimum	Maximum
A: Establishment Type and Age Range					
Population					
All Valid QWI Ages, All Establishments	2,957	1.027E-05	5.891E-06	2.149E-06	5.356E-05
All Valid QWI Ages, Private Establishments	2,957	1.152E-05	6.592E-06	1.990E-06	5.403E-05
B: State					
Postal Code					
AK	188	5.625E-05	4.128E-05	7.485E-06	1.601E-04
AL	172	2.873E-05	2.214E-05	3.299E-06	8.659E-05
AR	148	3.144E-05	2.094E-05	6.305E-06	7.746E-05
AZ	124	4.139E-05	3.343E-05	5.211E-06	1.432E-04
CA	324	2.273E-05	1.914E-05	2.422E-06	8.591E-05
CT	252	3.258E-05	2.605E-05	5.610E-06	1.371E-04
DC	104	8.303E-05	5.866E-05	1.708E-05	2.181E-04
DE	212	6.386E-05	5.424E-05	5.393E-06	2.005E-04
FL	304	1.645E-05	1.296E-05	2.559E-06	6.220E-05
GA	220	2.916E-05	2.281E-05	3.254E-06	8.538E-05
HI	256	4.090E-05	4.112E-05	5.684E-06	2.272E-04
IA	208	2.801E-05	2.095E-05	4.748E-06	7.924E-05
ID	332	6.090E-05	4.640E-05	8.729E-06	1.853E-04
IL	348	2.404E-05	1.901E-05	2.408E-06	6.948E-05
IN	220	2.254E-05	1.722E-05	3.352E-06	6.253E-05
KS	300	3.776E-05	3.054E-05	5.557E-06	1.261E-04
KY	172	2.588E-05	1.884E-05	3.681E-06	8.234E-05
LA	268	2.087E-05	1.429E-05	3.630E-06	5.697E-05
MD	348	2.926E-05	2.510E-05	3.640E-06	1.148E-04
ME	248	2.783E-05	1.945E-05	4.771E-06	8.563E-05
MI	180	1.626E-05	1.206E-05	2.406E-06	4.997E-05
MN	276	2.264E-05	1.861E-05	2.661E-06	6.325E-05
MO	268	2.164E-05	1.737E-05	3.001E-06	6.110E-05
MS	132	3.374E-05	2.393E-05	4.847E-06	9.564E-05
MT	300	4.097E-05	2.925E-05	7.754E-06	1.329E-04
ND	220	4.407E-05	3.356E-05	5.709E-06	1.191E-04
NE	204	3.838E-05	2.992E-05	4.032E-06	1.109E-04
NH	140	3.957E-05	2.731E-05	6.623E-06	9.685E-05
NJ	252	2.814E-05	2.202E-05	4.148E-06	9.915E-05
NM	260	6.823E-05	5.648E-05	6.973E-06	2.258E-04
NV	220	5.935E-05	4.752E-05	6.652E-06	1.766E-04
NY	188	2.177E-05	1.894E-05	2.341E-06	9.451E-05
OH	188	1.402E-05	1.088E-05	2.160E-06	3.793E-05
OK	188	4.342E-05	3.421E-05	3.832E-06	1.146E-04
OR	332	3.542E-05	2.710E-05	5.739E-06	1.056E-04
PA	236	1.094E-05	7.845E-06	1.990E-06	3.553E-05
RI	268	5.713E-05	4.069E-05	9.523E-06	1.553E-04
SC	220	3.390E-05	2.587E-05	4.492E-06	1.016E-04
SD	220	4.917E-05	3.734E-05	7.047E-06	1.459E-04
TN	220	2.221E-05	1.741E-05	2.674E-06	6.856E-05
TX	268	1.757E-05	1.413E-05	2.003E-06	6.805E-05
UT	196	6.683E-05	5.510E-05	7.998E-06	1.977E-04
VA	220	2.636E-05	2.214E-05	3.717E-06	1.089E-04
VT	188	4.051E-05	2.909E-05	7.829E-06	1.249E-04
WA	348	2.724E-05	2.122E-05	3.714E-06	7.521E-05
WI	348	2.052E-05	1.673E-05	3.026E-06	6.773E-05
WV	236	2.199E-05	1.394E-05	3.598E-06	6.286E-05
WY	172	8.041E-05	6.154E-05	1.300E-05	2.730E-04

Notes: There is small amount of between-implicate variance of state counts for full-quarter employment. We summarize the between variance using the coefficient of variation defined as the square root of the between-implicate variance divided by the average between-implicate weighted counts. Panel A summarizes the coefficient of variation for the between variance for the four different types of ownership type and age populations. The summary is taken across all state-year-quarters. Panel B summarizes the coefficient of variation for all states across all year, quarters, and ownership types and age range combinations.

BIBLIOGRAPHY

Aaronson, D. and E. French (2004). The effect of part-time work on wages: Evidence from the social security rules. *Journal of Labor Economics 22*(2), 329–352.

Aaronson, S. and A. Figura (2010). How biased are measures of cyclical movements in productivity and hours? *Review of Income and Wealth 56*(3), 539–558.

Abowd, J. M., K. Gittings, K. L. McKinney, B. E. Stephens, L. Vilhuber, and S. Woodcock (2012). Dynamically consistent noise infusion and partially synthetic data as confidentiality protection measures for related time-series. In *Federal Committee on Statistical Methodology, 2012 Research Conference Papers*. Office of Management and Budget.

Abowd, J. M., B. E. Stephens, L. Vilhuber, F. Andersson, K. L. McKinney, M. Roemer, and S. Woodcock (2009). The LEHD Infrastructure Files and the Creation of the Quarterly Workforce Indicators. In T. Dunne, J. B. Jensen, and M. J. Roberts (Eds.), *Producer Dyanamics: New Evidence from Micro Data*, pp. 149–230. University of Chicago Press.

Abowd, J. M. and M. H. Stinson (2013). Estimating Measurement Error in Annual Job Earnings: A Comparison of Survey and Administrative Data. *Review of Economics and Statistics* (August), —-.

Abowd, J. M. and L. Vilhuber (2005). The Sensitivity of Economic Statistics to Coding Errors in Personal Identifiers. *Journal of Business & Economic Statistics 23*(2), 133–152.

Abraham, K. G., J. Haltiwanger, K. Sandusky, and J. R. Spletzer (2013). Exploring Differences in Employment between Household and Establishment Data. *Journal of Labor Economics 31*(S1), S129–S172.

Altonji, J. G. and C. H. Paxson (1986). Job characteristics and hours of work. *Research in Labor Economics 8A*, 1–55.

Altonji, J. G. and C. H. Paxson (1988). Labor supply preferences, hours constraints, and hours-wage trade-offs. *Journal of Labor Economics 6*(2), 254–276.

Ashenfelter, O. and R. S. Smith (1979). Compliance with the Minimum Wage Law. *Journal of Political Economy 87*(2), 333–350.

Autor, D. H. and M. G. Duggan (2003). The rise in the disability rolls and the decline in unemployment. *Quarterly Journal of Economics 118*(1), 157–206.

Bartik, T. J. (1991). *Who Benefits from State and Local Economic Development Policies?* W.E. Upjohn Institute for Employment Research.

Basu, A. K., N. H. Chau, and R. Kanbur (2010). Turning a Blind Eye: Costly Enforcement, Credible Commitment and Minimum Wage Laws. *The Economic Journal 120*(543), 244–269.

Baum-Snow, N. and D. Neal (2009). Mismeasurement of usual hours worked in the census and ACS. *Economics Letters 102*(1), 39–41.

Benedetto, G., J. Haltiwanger, J. Lane, and K. Mckinney (2007). Using Worker Flows to Measure Firm Dynamics. *Journal of Business & Economic Statistics 25*(3), 299–313.

Bernhardt, A., M. Spiller, and N. Theodore (2013). Employers gone rogue: Explaining industry variation in violations of workplace laws. *Industrial and Labor Relations Review 66*(4), 808–832.

Biddle, J. E. (2014). Retrospectives: The Cyclical Behavior of Labor Productivity and the Emergence of the Labor Hoarding Concept. *Journal of Economic Perspectives 28*(2), 197–212.

Bishop, Y. M., S. E. Fienberg, and P. W. Holland (1975). *Discrete Multivariate Analysis: Theory and Practice*. Cambridge, MA: MIT Press.

Blanchard, O. J. and F. Katz (1991). Regional Revolutions. *Brookings Papers on Economic Activity* (1), 1–75.

Borowczyk-Martins, D. and E. Lalé (2015). How bad is involuntary part-time work? *mimeo*.

Bound, J., C. Brown, G. J. Duncan, and W. L. Rodgers (1994). Evidence on the Validity of Cross-Sectional and Longitudinal Labor Market Data. *Journal of Labor Economics 12*(3), 345.

Bound, J., C. Brown, and N. Mathiowetz (2001). Measurement error in survey data. In J. Heckman and E. Leamer (Eds.), *Handbook of Econometrics*, Volume 5, Chapter 59, pp. 3705–3843. Elsevier Masson SAS.

Bound, J. and H. J. Holzer (2000). Demand Shifts, Population Adjustments, and Labor Market Outcomes during the 1980s. *Journal of Labor Economics 18*(1), 20–54.

Bound, J. and A. B. Krueger (1991). The Extent of Measurement Error in Longitudinal Earnings Data: Do Two Wrongs Make a Right? *Journal of Labor Economics 9*(1), 1.

Burda, M., K. R. Genadek, and D. S. Hammermesh (2016). Not Working at Work. *NBER Working Paper Series 21923*.

Burgess, M. (2014). How frequently do private businesses pay workers? *Bureau of Labor Statistics: Beyond the Numbers 3*(11).

Burgess, S., J. Lane, and D. Stevens (2000). Job flows, worker flows, and churning. *Journal of Labor Economics 18*(3), 473–502.

Caballero, B. R. J., E. M. R. A. Engel, and J. Haltiwanger (1997). American Economic Association Aggregate Employment Dynamics : Building from Microeconomic Evidence. *The American Economic Review 87*(1), 115–137.

Caballero, R. J. and E. M. R. A. Engel (1993). Microeconomic Adjustment Hazards and Aggregate Dynamics. *The Quarterly Journal of Economics 108*(2), 359–383.

Cajner, T., D. Mawhirter, C. Nekarda, and D. Ratner (2014). Why is involuntary part-time work elevated? *FEDS Notes*.

Cameron, A. C. and D. L. Miller (2015). A Practitioner's Guide to Cluster- Robust Inference. *Journal of Human Resouces 50*(2), 317–372.

Canon, M., M. Kudlyak, and M. Reed (2014). Is involuntary part-time employmet different after the great recession? *Federal Reserve Bank of St. Louis: The Regional Economist*.

Chang, Y.-m. and I. Ehrlich (1985). On the Economics of Compliance with the Minimum Wage Law. *Journal of Political Economy 93*(1), 84–91.

Cooper, R., J. Haltiwanger, and J. L. Willis (2004, February). Dynamics of labor demand: Evidence from plant-level observations and aggregate implications. *NBER Working Paper 10297*.

Cooper, R., J. Haltiwanger, and J. L. Willis (2007). Search frictions: Matching aggregate and establishment observations. *Journal of Monetary Economics 54*(SUPPL.), 56–78.

Cooper, R. and J. L. Willis (2009). The cost of labor adjustment: Inferences from the gap. *Review of Economic Dynamics 12*, 632–647.

Davis, S., R. J. Faberman, and J. Haltiwanger (2012). Labor market flows in the cross section and over time. *Journal of Monetary Economics 59*(1), 1–18.

Davis, S. J. and J. Haltiwanger (1990). Gross job creation and destruction: Microeconomic evidence and macroeconomic implications. *NBER Macroeconomics Annual 5*, 123–168.

Davis, S. J. and J. Haltiwanger (1999). Gross job flows. *Handbook of Labor Economics: eds. O. Ashenfelter and D. Card 3B*, 2711–2805.

Duncan, G. J. and D. H. Hill (1985). An Investigation of the Extent and Consequences of Measurement Error in Labor-Economic Survey Data. *Journal of Labor Economics 3*(4), 508.

Eldridge, L. P. and S. W. Pabilonia (2010). Bringing work home: implications for {BLS} productivity measures, Monthly Labor Review Online, December 2010. *Monthly Labor Review 133*(12), 18–35.

Evans, T., L. Zayatz, and J. Slanta (1998). Using Noise for Disclosure Limitation of Establishment Tabular Data. *Journal of Official Statistics 14*(4), 537–551.

Even, W. E. and D. A. Macpherson (2015). The affordable care act and the growth of involuntary part-time employment. *IZA Discussion Paper Series No. 9324*.

Fair, R. C. (1969). *The Short-Run Demand for Workers and Hours*. Amsterdam: North-Holland Publishing Company.

Farber, H. S. (1999a). Alternative and part-time employment arrangements as a response to job loss. *Journal of Labor Economics 17*(4), S142–S169.

Farber, H. S. (1999b). Mobility and stability: The dynamics of job change in labor markets. *Handbook of Labor Economics: eds. O. Ashenfelter and D. Card 3B*, 2439–2483.

Frazis, H. and J. Stewart (2004). What can time-use data tell us about hours of work. *Monthly Labor Review.* (December), 3–9.

Frazis, H. and J. Stewart (2009). Comparing Hours per Job in the CPS and the ATUS. *Social Indicators Research 93*(1), 191–195.

Frazis, H. and J. Stewart (2010). Why Do BLS Hours Series Tell Different Stories about Trends in Hours Worked? In K. G. Abraham, J. R. Spletzer, and M. Harper (Eds.), *Labor in the New Economy*, Number October, pp. 343–372. University of Chicago Press.

Hall, R. (2004). Measuring factor adjustment costs. *Quarterly Journal of Economics 119*(3), 899–927.

Haltiwanger, J. C., S. J. Davis, and S. Shuh (1996). *Job Creation and Job Destruction*. Cambridge, MA: MIT Press.

Hamermesh, D. S. (1989a). Labor Demand and the Structure of Adjustment Costs. *The American Economic Review 79*(4), 674–689.

Hamermesh, D. S. (1989b). Labor demand the structure of adjustment costs. *American Economic Review 79*(4), 674–689.

Hamermesh, D. S. and G. A. Pfann (1996). Adjustment costs in factor demand. *Journal of Economic Literature 34*(3), 1264–1292.

Higgins, C., L. Duxbury, and K. L. Johnson (2000). Part-time work for women: Does it really help balance work and family? *Human Resource Management 39*(1), 17–32.

Hijzen, A. and D. Venn (2011). The role of short-time work schemes during the 2008-09 recession. *OECD Social, Employment and Migration Working Papers No. 115*.

Ji, M. and D. Weil (2015). The Impact of Franchising on Labor Standards Compliance. *ILR Review 68*(5), 977–1006.

Kalecki, M. (1943). Political Aspects of Full Employment. *Political Quarterly 14*, 322–330.

Kalleberg, A. L. (2000). Nonstandard employment relations: Part-time, temporary, and contract work. *Annual Review of Sociology 26*, 341–365.

Lazear, E. P., K. L. Shaw, and C. Stanton (2015). Making Do with Less: Working Harder during Recessions. *Journal of Labor Economics 34*(1), 1–44.

Lazear, E. P. and J. R. Spletzer (2012). Hiring, churn, and the business cycle. *American Economic Review: Papers and Proceedings 102*(3), 575–579.

Lee, Y. and T. Mukoyama (2012). Entry, exit, and plant-level dynamics over the business cycle. *Federal Reserve Bank of Cleveland Working Paper No. 07-18.*.

Lettau, M. K. (1997). Compensation in part-time jobs versus full-time jobs: What if the job is the same? *Economics Letters 56*(1), 101–106.

Li, Q. and J. Racine (2003). Nonparametric Estimation of Distributions with Categorical and Continuous Data. *Journal of Multivariate Analysis 86*(2), 266–292.

Little, R. J. and D. B. Rubin (2002). *Statistical Analysis with Missing Data* (2nd ed.). Hoboken, NJ: Wiley.

Mellow, W. and H. Sider (1983). Accuracy of Response in Labor Market Surveys: Evidence and Implications. *Journal of Labor Economics 1*(4), 331.

Mendeloff, J., C. Nelson, K. Ko, and A. Haviland (2006). Small Businesses and Workplace Fatality Risk: An exploratory analysis. Technical report, Rand Corporation.

Meyer, B. D. and R. M. Goerge (2011). Errors in Survey Reporting and Imputation and their Effects on Estimates of Food Stamp Program Participation.

Milkman, R., A. L. González, and P. Ikeler (2012). Wage and hour violations in urban labour markets: a comparison of Los Angeles, New York and Chicago. *Industrial Relations Journal 43*(5), 378–398.

Oi, W. Y. (1962). Labor as a Quasi-Fiexed Factor. *Journal of Political Economy 70*(6), 538–555.

Rebitzer, J. B. (1987). Unemployment, Long-Term Employment Relations, and Productivity Growth. *The Review of Economics and Statistics 69*(4), 627–635.

Rubin, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys*. New York: John Wiley & Sons.

Rubin, D. B. and N. Schenker (1986). Multiple Imputation for Interval Estimation From Simple Random Samples With Ignorable Nonresponse. *Journal of the American Statistical Association 81*(394), 366–374.

Schaller, J. S. (2016). Booms , Busts , and Fertility: Testing the Becker Model Using Gender-Specific Labor Demand. *The Journal of Human Resources 51*(1), 1–29.

Shapiro, C. and J. E. Stiglitz (1984). Equilibrium Unemployment as a Worker Descipline Device. *American Economic Review 74*(3), 433–444.

Shimer, R. (2012). Reassessing the ins and outs of unemployment. *Review of Economic Dynamics 15*(2), 127–148.

Starsinic, M. (2011). Incorporating a Finite Population Correction Factor into American Community Survey Estimates. In *JSM Proceedings, Survey Research Methods Section*, Alexandria, Virginia, pp. 3621–3631. American Statistical Association.

Tibshirani, R. (1996). Regression Selection and Shrinkage via the Lasso. *Journal of the Royal Statistical Society B 58*(1), 267–288.

Topel, R. H. and M. P. Ward (1992). Job mobility and the careers of young men. *The Quarterly Journal of Economics 107*(2), 439–479.

Tornqvist, L., P. Vartia, and Y. O. Vartia (1985). How Should Relative Changes Be Measured? *The American Statistician 39*(1), 43–46.

U.S. Bureau of Labor Statistics (2004). Construction of Average Weekly Hours for Supervisory and Nonproduction Wage and Salary Workers in Private Nonfarm Establishments. Technical report, Bureau of Labor Statistics.

U.S. Census Bureau (2003a). 2000 Census of Population and Housing, Public Use Microdata Sample, United States: Technical Documentaiton.

U.S. Census Bureau (2003b). PUMS Accuracy of the Data, The American Community Survey.

U.S. Census Bureau (2015). American Community Survey Multiyear Accuracy of the Data.

Valletta, R. and L. Bengali (2013). What's behind the increase in part-time work? *FRBSF Economic Letter 2013-24*.

Valletta, R., L. Bengali, and C. van der List (2016). Cyclical and market determinants of involuntary part-time employment. *IZA Discussion Paper Series No. 9738*.

Valletta, R. and C. van der List (2015). Involuntary part-time work: Here to stay? *FRBSF Economic Letter 2015-19*.

Wagner, D. and M. Layne (2014). The Person Identificatoin Validation System (PVS): Applying the Center for Administrative Records Research and Applications' (CARRA) Record Linkage Software.

Weil, D. (1991). Enforcing OSHA: The Role of Labor Unions. *Industrial Relations 30*(1), 20–36.

Weil, D. (2010). Improving Workplace Conditions Through Strategic Enforcement: A Report to the Wage and Hour Division. Technical report.

Zabalza, A., C. Pissarides, and M. Barton (1980). Social security and the choice between full-time work, part-time work and retirement. *Journal of Public Economics 14*(2), 245–276.

Zhang, X., M. L. King, and R. J. Hyndman (2006). A Bayesian Approach to Bandwidth Selection for Multivariate Kernel Density Estimation. *Computational Statistics & Data Analysis 50*(11), 3009 – 3031.